

Advisers between a rock and a hard place

The possibility that bovine spongiform encephalopathy (BSE) may have spread to sheep is causing concern to scientific advisory committees in Europe. More research is urgently required, but so too is more openness.

Serving on scientific committees that advise governments can be a privilege. But there are circumstances where the judgments of experts can have far reaching and important consequences; where risks seem remote, but where the outcome were the worst to occur could be damaging to the economic and public health of nations. A government may then be highly dependent on scientific advice, and the experts find themselves in an utterly unenviable position.

Have some sympathy, therefore, for the members of the Spongiform Encephalopathy Advisory Committee (SEAC), which advises the British ministries of agriculture and of health, and its equivalents elsewhere in Europe. For the quandary they face is that BSE, which is suspected of having passed from cows to create a new variant of Creutzfeldt-Jakob disease in humans (vCJD), may also be lurking in sheep flocks. Given that sheep were certainly exposed to BSE-contaminated meat and bonemeal, experts reckon that the risk is real that some BSE may be masquerading as the sheep's equivalent of BSE, scrapie (see page 6).

Inconclusive

So far there is no conclusive evidence that this has happened. Indeed, the experimental evidence either way is slim. In one set of experiments, a Cheviot sheep inoculated with infected cattle brain succumbed to the disease, the agent adopting the pattern of infection of scrapie — spreading wider within the body than it does in cows. The experiments have not been repeated or extended to other breeds. Experiments to test whether sheep can contract BSE from contaminated feed have also not been done.

One optimistic view is that, even if BSE has entered the sheep population, it may have left it again, following the 1988 ban on meat and bonemeal feeds. But, without more data, experts have been forced to conclude that in theory sheep BSE could represent the worst of both worlds — having the killer pathology of cow BSE, but spreading from animal to animal as scrapie does. As a precaution, the United Kingdom and France have therefore banned sheep heads, spinal cords and spleen from the human and animal food chains.

Unfortunately, distinguishing scrapie from BSE in a single sheep is itself full of uncertainties, very costly, and can take over a year to accomplish. Only nine sheep have been analysed so far and, while none was found to have BSE, that is scant cause for comfort, given that there are an estimated 100 million sheep in Europe alone. To make matters worse, there are major gaps and uncertainties in the epidemiology of scrapie, in the genetic characteristics of susceptibility to it, and in the histories of sheep feeding and rendering practices across Europe. Knowledge of the mechanism underlying these diseases is also flimsy.

Where the uncertainties are so pervasive, further measures to reduce risk could only be correspondingly broad in scope and huge in cost and misery. A ban on sheep offals or all mutton, for example,

would seem untenable at present, in particular given that only 27 cases of vCJD have been confirmed so far — although it is still too early to predict the scale of any vCJD epidemic.

SEAC and its equivalents in Brussels and elsewhere can do little more than the obvious: spell out the limited facts and the uncertainties, and recommend spending considerable sums of money on extensive and urgent research — a relatively small and wise investment compared to retrospective action, considering that the BSE crisis has already cost the European Union some ECU3.5 billion (US\$3.9 billion) in beef subsidies.

Transparency

If, as is to be hoped, the chances of BSE having spread to sheep are negligible, it may take years to demonstrate the fact; if it has, such drastic measures would become a political possibility. Only politicians can take responsibility for future or present precautions — but the interplay with experts and the public is crucial in determining the difficulties and impact of such decisions.

The readiness of society to understand and anticipate precautions depends in turn on the openness of debate on risk and the scientific advice given. Openness has drawbacks. Complex issues may be misunderstood or misrepresented. Moreover, genuine damage may be done to those who, in all innocence, lose some of their livelihood or reputation through the enhanced perception of risk.

But those penalties must be weighed against the costs of secrecy: that the science itself is undermined if not discussed openly, and that a public which suspects it has been kept in the dark and thereby misled may overreact to the sudden announcement of profoundly bad news. Not least, there is the fundamental issue of the rights of a citizen in a democracy — SEAC's only consumer representative makes the valid point that, although she is not a scientist, the access to data which she enjoys as a member of SEAC allows her to make an informed judgement on what she eats that is largely denied to the rest of the public.

In his 1997 note, *The use of scientific advice in policy making*, Sir Robert May, chief scientific adviser to the UK government, insisted on a presumption towards openness: "departments should aim to publish all the scientific evidence and analysis underlying policy decisions on the sensitive issues — an action which could in itself avoid greater controversy in the longer run".

Sir Robert is right. SEAC would be doing itself and the public a favour by making all the information readily available in the public domain in comprehensible language, organizing a public consensus conference on the issue, and creating a comprehensive SEAC website.

Fears that such openness might create public hysteria should give rise to sensitivity rather than secrecy. If there is one lesson that governments are learning, it is surely that panic feeds on ignorance and a lack of anticipation. □



Japan's science agency seeks extra billions, despite recession

[TOKYO] Japan's Science and Technology Agency is planning a dramatic increase in its budget for brain science and genomics next fiscal year. The agency's budget request for 1999 submitted on Monday this week (31 August) also includes new funds for a deep-sea drilling ship that is intended to play a key role in the extension of the international Ocean Drilling Programme.

Despite Japan's economic woes, the government continues to plough new money into science and technology. The Science and Technology Agency is asking for an increase of only 2.9 per cent in its standard budget for fiscal year 1999, which starts in April next year. But it is also requesting an extra ¥17.8 billion (US\$123 million), equivalent to an additional 2.3 per cent, under a special appropriation to develop infrastructure for telecommunications, science and technology, and the environment for the next century.

This special one-off appropriation, which is available for all science-related ministries and agencies in 1999 and is expected to total ¥150 billion, is intended to help Japan establish a sounder economic base. The agency's request may be cut back before approval by the Cabinet and Ministry of Finance at the end of the year, but such adjustments are usually minor.

The agency is requesting ¥9.1 billion — ¥6.8 billion more than this year — to bring together the life sciences and information sciences in a broad new area of research that encompasses bioinformatics but also covers elucidation of such things as information processing in the brain. Some ¥6.4 billion will come from the special appropriation for this programme.

Partly because of this appropriation, the budget for genome science is increased 66 per cent to ¥11.6 billion, including ¥7.7 billion for a new genomic science centre to be established by the Institute of Physical and Chemical Research (RIKEN) in Yokohama (see *Nature* 392, 219; 1998). Some ¥3.3 billion of the total will come from the special appropriation.

The centre will open a temporary office in October at RIKEN's main campus in Wako city until its new building is completed in Yokohama in 2000. It will focus on three areas of research: cDNA sequencing and investigation of genome function for both humans and mice; development of a 'gene encyclopedia' of the mouse genome; and an NMR park for investigating protein function and structure.

Similarly, the budget for brain science is increased 36 per cent to ¥19 billion, includ-

	Request (¥ billion)	Change since 1998 (%)
Total budget	761.8*	+2.9
Special appropriation for 21st century	17.8	—
Life/information sciences	9.1	+386
Genome science	11.6	+66
Brain research	19.0	+36
Deep-sea drilling ship	5.7	+946

* Excluding special appropriation for 21st century.

ing ¥11.7 billion for the new Brain Science Institute at RIKEN's main campus. The increased budget also includes ¥3 billion of the new money for life and information sciences.

Other beneficiaries of the special appropriation will be the world's marine geologists. The agency is requesting ¥5.7 billion for a deep-sea drilling ship expected to be completed in 2003. Researchers at the Japan Marine Science and Technology Center, who

are leading the programme, are confident of getting a total budget of US\$120 million for the ship's design and construction over the next three years.

The ship will have 'riser technology' adapted from the petroleum industry and will allow drilling to much greater depths in the Earth's crust and in more difficult geological terrain. The ship is intended to play a key role in the future of the Ocean Drilling Programme beyond 2003, when the current phase of the programme ends (see *Nature* 388, 409; 1997 & 394, 820; 1998).

The agency's budget also gives the first signs of the pending merger of the agency with the Ministry of Education, Science, Sports and Culture (see *Nature* 390, 327; 1997). The ministry and agency are making joint requests for some projects, for example, to fund research on the Moon by both the National Space Development Agency, affiliated to the agency, and the Institute of Space and Astronautical Science under the education ministry.

David Swinbanks

Peregrine leads flight from endangered list

[WASHINGTON] The US interior department last week proposed removing the peregrine falcon from its list of endangered species. This would be the first time in three years that a species has been 'delisted' because of a dramatic recovery in numbers.

The peregrine is the first of about two dozen species scheduled to be taken off the list in the next two years, following a directive in May by interior secretary Bruce Babbitt to expedite delistings.

More than 1,100 animals and plants are protected under the Endangered Species Act (ESA). But only 27 species, subspecies or populations have been taken off the list since the act came into force in 1973.

John Fay, a biologist at the endangered species division of the US Fish and Wildlife Service, says: "We're pretty conservative about calling something recovered." Because some endangered populations had been declining for as long as a century before the ESA was enacted, he says, "after 25 years it's unreasonable to expect a lot of recoveries".

The service has lacked the resources to devote to delisting, having only recently recovered from a congressional moratorium on new listings in 1995 that created a backlog of species in need of protection.

The peregrine has bounced back from a low of 324 nesting pairs in 1975 to at least 1,593 breeding pairs in the United States and



Out of danger: US peregrines stage a comeback.

Canada. The comeback is due more to the banning of the pesticide DDT in the 1970s and to an aggressive captive breeding programme than to ESA protection. This has led congressional critics to attack Babbitt for touting the peregrine and other delisting candidates as ESA success stories.

Last month, 30 members of the Western Caucus, a conservative group of property rights advocates in the House of Representatives, wrote to Babbitt challenging his statement in May that the ESA has been proved a success. Some species are being removed from the list, they point out, only because they have become extinct or have been reclassified, and others are being delisted because their population estimates have grown as scientists discover more of them in the wild. Most biologists and wildlife managers support delisting in principle, but urge caution. **Tony Reichhardt**

NIAL BENVIE/BBC NAT. HIST. UNIT

NSF director backs environment studies

[WASHINGTON] The National Science Foundation (NSF) is planning a major environmental science initiative that will satisfy pressure from some Congressmen for it to accommodate a 'national institute for the environment' (NIE), while reflecting the interests of Rita Colwell, the marine biologist who was sworn in last month as the foundation's new director.

Earlier this year, the National Science Board, which oversees the NSF, asked Jane Lubchenco of Oregon State University to chair a task force on the environment that will report back early in 1999. The board also emphatically rejected the NIE proposal, saying it "would create an additional and wasteful structure and bureaucracy where no such need exists".

Now Colwell, a former proponent of an NIE, predicts that the study will lead to a new initiative in environmental research. "We are

now beginning to invest in the early stages of a biocomplexity/ecosystem dynamics initiative, and we are addressing this in a very serious way," she said in an interview.

Colwell was a member of the board of directors of the Committee for the National Institute of the Environment, which has been campaigning in Washington for several years to create a new federal agency to perform environmental research. She resigned from the board in February, when she was named by President Bill Clinton to head the NSF. But she denies having changed her position on the need for such an agency. "The call was for increased research in the environment," she says. "I was never a proponent of a separate, independent agency, but I remain an ardent proponent of more environmental research."

Colwell says that her priorities for the agency, which funds most non-biomedical



Colwell: outlining agency's priorities.

university research in the United States, would include interdisciplinary research, science education and international scientific collaboration.

She says she wants the NSF to be at the forefront of advanced computing research. Last week she met Ernest Moniz, the under-secretary of energy, to discuss possible NSF involvement in Moniz's plan for a major supercomputing initiative at the Department of Energy's non-weapons laboratories (see *Nature* 390, 651; 1997).

Energy department sources and Colwell say that the initiative is likely to be in the Clinton administration's budget request for the year 2000 as a multi-agency initiative involving the Department of Energy, the NSF, the National Institutes of Health and other science agencies.

Colwell says the NSF — which is administratively divided into directorates and divisions that address specific scientific disciplines — is already improving its mechanisms for supporting interdisciplinary work. She points, for example, to its recent establishment of an office of integrated activities, which has a dozen staff. The office is "an indication that we're moving in some new directions", she says. Pointing out that the NSF structure reflects the structure of the universities, which "is undergoing major change", Colwell declines to rule out a major reorganization of the agency. "I've only been here three weeks — I need time to work with the associate directors," she says.

Colwell, who was director of the multidisciplinary University of Maryland Biotechnology Institute at Baltimore, also defends her role in winning federal funds for the Christopher Columbus Center, a \$160 million research and education centre in Baltimore, whose public outreach activities have been temporarily shut down after running into financial difficulties. "Every single research project in that building has been peer-reviewed", she says. "The one part that hasn't done too well wasn't under my control at all. It is being reorganized and it will be a success."

She adds that the NSF is currently examining its own administrative needs, as its staff has been held at 1,200 for ten years while its budget has doubled to \$3.5 billion (see *Nature* 393, 100; 1998). "We're definitely looking into it. Joe [Bordogna, acting deputy director of NSF] has already initiated some studies on the organizational structure here."

Colin Macilwain

US drive to speed trials for rare disorders

[WASHINGTON] The US Food and Drug Administration (FDA) and the National Institutes of Health (NIH) are launching an initiative to make clinical trials for rare illnesses speedier, less cumbersome and less costly.

The FDA and NIH aim to promote research on gene therapies for rare disorders by luring more investigators into the area and encouraging industry to support research in clinical areas that do not have a big market.

The NIH has received 259 applications for gene therapy protocols in the past ten years. But few involve rare diseases, even though gene therapy was conceived to treat rare, inherited disorders due to mutations in a single gene. These 'monogenic' diseases are increasingly ignored as scientists and their industry sponsors focus on developing gene therapies for large-market diseases such as AIDS and cancer.

Only 34 protocols have been for monogenic diseases. By contrast, 61 per cent have been for cancer. And the trend is growing

more marked. Of 31 protocols submitted so far in 1998, two are for monogenic diseases while 22 are for cancer.

"All the money is going into the most profitable diseases... so it's obviously an area that the government has to step into," says Abbey Meyers, the president of the National Organization for Rare Disorders.

Philip Noguchi, the director of FDA's Division of Cellular and Gene Therapies, says gene therapy's "focus [on rare diseases] has drifted over the last few years. This [new initiative] is an attempt to bring it back."

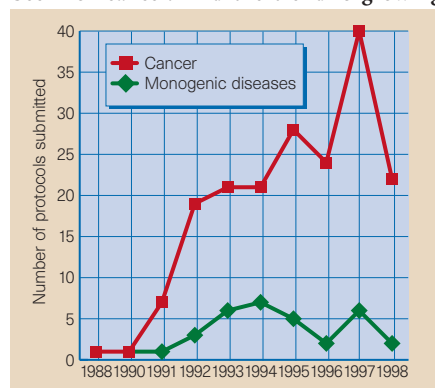
A working group from the FDA's Center for Biologics Evaluation and Research and the NIH's Office of Rare Diseases is developing a strategy to speed up trials.

First, it wants to relieve investigators of the need to reinvent the wheel each time they apply to do a gene therapy experiment. Currently, each investigator must prove a given vector is safe for a proposed therapy. The group hopes to demonstrate through NIH-funded studies that a particular vector is broadly safe so investigators will not have to demonstrate its safety separately.

Second, the group proposes that the FDA accepts protocols with as few patients as possible and merges safety and efficacy studies, which are conducted in separate stages.

Even if the initiative greatly reduces the early development costs for companies, however, there remains little incentive to commercialize for small target populations. James Wilson, the director of the Institute for Human Gene Therapy at the University of Pennsylvania, says he is "afraid we will still be left with some diseases in which we can't encourage anybody to do it".

Meredith Wadman



Protocols submitted to the NIH Recombinant DNA Advisory Committee, 1988–98.

Royal Society wants genetics watchdog

[LONDON] The Royal Society has called on the British government to set up a regulatory body to oversee all aspects of developing genetically modified food. The society also wants an end to the use of antibiotic-resistance marker genes in genetically modified food products.

The recommendations are made in a report assessing the science and regulation of genetically modified food products. Released today (3 September), the report was written by an expert group set up by the society to address the current public controversy over genetically modified food.

The report is broadly enthusiastic about the potential of genetic modification to address future food requirements. But it voices concerns about the lack of coordination of regulations in Britain, and says, as had been expected (see *Nature* 394, 207; 1998), that there should be a single, "over-arching" body to monitor all safety and regulatory aspects of genetic engineering in agriculture and food production.

Companies wishing to grow or sell genetically modified produce must currently comply with a stepwise set of regulations governing laboratory studies, research and com-

mercial field trials, and entry into the food chain. Regulations are set and monitored by different government departments, which are in turn advised by their respective scientific advisory committees.

The proposed body would oversee the enforcement of these regulations, keep track of the whole "life history" of a genetically modified crop plant, and monitor research questions such as the transfer of genes between crops. The report urges the government "not to delay further action in taking this direction".

Public concerns about genetic modification are also covered in the report. However, it does not advocate a labelling scheme for all food containing genetically modified organisms. In line with European Union regulations, it "strongly supports" labelling "where the new foodstuff is substantially changed from that of its conventional counterpart".

The report recognizes the difficulties faced by retailers who want to buy non-genetically modified food where suppliers do not segregate modified and non-modified varieties. It says research is needed to develop "scientifically validated testing methods" that can recognize an agreed minimum level

of proteins produced by modified genes.

It also draws up a list of issues on genetic modification that need further research. These include using alternatives to antibiotic-resistance marker genes and assessing the impact of virus-resistant and insect-tolerant plants on ecosystems.

However, the report does not support a moratorium on the commercial release of genetically modified crops. A moratorium pending the outcome of research issues is advocated by community pressure groups and environmentalist and consumer organizations, as well as by the government's wildlife advisory body, English Nature.

The society says it considers the risks of genes spreading from genetically modified crop plants to wild species to be slight. It adds that there is little evidence from field and laboratory studies that such crops have an adverse impact on other plants or on insects.

The report does, however, acknowledge the scarcity of research data on the environmental impact of genetically modified herbicide-tolerant crops, as such crops have been grown commercially only since 1995 (in the United States and Canada, but not in the United Kingdom).

Ehsan Masood

China brings in regulations to put a stop to 'genetic piracy'

[SHANGHAI] China has adopted a broad set of regulations on the collection and use of its 'human genetic resources' in an attempt to restrict their exploitation by foreign biotechnology and pharmaceutical companies.

China is becoming increasingly aware that its large population and relatively high number of distinct ethnic groups provide fertile ground for the search for disease genes that could hold the key to profitable diagnostic and therapeutic techniques.

Foreign companies have reportedly been obtaining data from studies carried out in China in collaboration with local researchers without obtaining official approval or ensuring any significant return to Chinese research institutions.

But, under the new regulations, official authorization will be required for any research project that seeks to "sample, collect, merchandize or export" human genetic resources. These are defined as "any materials of and from human beings that contain human genome, genes or gene products, or parts thereof".

The regulations also specify that, if valuable genetic information emerges from a collaboration between a Chinese research institution and a foreign organization, profits from the resultant patents should be shared in proportion to the contribution of

the two bodies concerned. Requests for the approval of individual projects and for the export of genetic material will be examined and issued on a quarterly basis.

Collaborative projects already established will also be required to apply for approval.

Those who helped draft the regulations, which will be administered through a new Administration Office of National Human Genetic Resources of China, say they tried to ensure that the regulations are not too rigid or restrictive on collaborative projects.

Boqin Qiang, for example, first vice-president of the Chinese Academy of Medical Sciences and leader of the Beijing Human Genome Research Centre, emphasizes that Chinese scientists are keen to establish such projects with foreign research groups, and that the regulations are not intend to act as a barrier.

The regulations also require that proper informed consent should be obtained from all those who provide samples of genetic material for researchers, or, where appropriate, from their family members. This will bring Chinese research procedures in line with those in Western countries. "Obtaining informed consent in this way is an important part of the process," says Zhu Chen, director of the Shanghai Human Genome Research Centre, who has also been closely involved in drafting the regulations.



Chen says he strongly supports clauses in the regulations on the equitable sharing of intellectual property rights. "If people in China do some of the work involved in collecting and processing samples, then they should share in the commercial benefits in recognition of their contribution," he says.

In addition to the genome centres in Beijing and Shanghai (see *Nature* 394, 109; 1998), China currently has about 30 laboratories involved in human genome research. Work is under way to study the genetic basis in the Chinese population of various diseases, including cancers, leukaemia, schizophrenia and cardiovascular diseases.

David Dickson

Doubts over ability to monitor risks of BSE spread to sheep

Declan Butler

Has 'mad cow disease' – BSE – infected sheep in the United Kingdom? Scientific advisers to the UK government and the European Commission are calling for research to assess the risks to be stepped up.

Scientists studying bovine spongiform encephalopathy agree on one thing: that there is a very real risk that BSE may have passed into flocks of sheep and goats in the United Kingdom and elsewhere.

UK sheep were fed BSE-contaminated meat and bone meal until the July 1988 feed ban, and limited experiments have shown that sheep fed infected cattle brain can succumb to BSE.

It remains unknown whether BSE has entered the UK sheep population, but is going undetected because its symptoms are indistinguishable from those of scrapie, a spongiform encephalopathy endemic in UK flocks, but which does not seem to transmit to humans.

However, if BSE has really passed into the sheep population, the consequences for human health could be extremely serious. "We could be facing a potential national emergency," says one member of the UK government's Spongiform Encephalopathy Advisory Committee (SEAC). Given that BSE seems capable of infecting humans via the oral route, SEAC judges that infected sheep tissues could likewise "pose a potential risk to human health if eaten".

SEAC, which reports to the agriculture ministry and the health department, confirmed last month that the total number of UK cases of new-variant CJD, the human version of BSE, is 27. Epidemiologists say it is still too soon to reach any conclusions about the scale of any epidemic.

Not enough research is being done to assess how significant the danger of sheep BSE might be, according to many scientists. Unease persists within SEAC which is to set up a sub-committee to draft proposals for a large expansion of research. The European Commission's independent Scientific Steering Committee is also scheduled to recommend more research at the European level.

This year the UK Ministry of Agriculture, Fisheries and Food (MAFF) is spending £2.8 million (US\$4.6 million) of its £12.7 million BSE research budget on sheep spongiform encephalopathies, and the Biotechnology and Biological Sciences Research Council has budgeted £3 million over the next three years.

In July 1996, SEAC recommended more research to establish levels of scrapie in sheep and to assess the risks of BSE. An abattoir survey of the incidence of scrapie in sheep brains, suggested by the committee, began in August 1997, while the MAFF imported more than 1,000 scrapie-free sheep from New Zealand at the beginning of this year for pathogenesis and transmission research.

Uncertainty

Sir John Pattison, SEAC's chairman, says the new subcommittee "will be the best way of developing specific recommendations". A MAFF spokesperson declined to comment on research needs before the subcommittee has delivered its recommendations, but said MAFF would follow SEAC's advice.

SEAC members admit that, as a result of the current insufficient effort to detect BSE in the sheep population, decisions on public health measures are based on uncertainty rather than science. Epidemiological data on scrapie-like cases is so poor throughout Europe that it is impossible to even tell whether the incidence has changed over the past decade.

Scientists are particularly concerned about the speed and scale of experiments to detect BSE strains in sheep. Conventional

strain typing studies to distinguish scrapie from BSE are burdensome and take years. So far, results from only nine UK sheep isolates have been published.

Jack Cunningham, until recently Britain's agriculture minister, has publicly interpreted these results as reassuring: "So far all the samples studied have revealed scrapie and not BSE. Scrapie is a disease endemic in the UK and France for over 200 years and which is to our knowledge, while analogous to BSE, harmless to man."

Many scientists are more cautious. "Having found zero out of nine, what confidence can we attach to the statement 'BSE is not present in sheep'?" asks Jeffrey Almond, a virologist and professor of microbiology at the University of Reading, who is a member of SEAC. "The answer is very little; 'absence of evidence' is often confused with 'evidence of absence'."

He points out that, to have a hope of detecting BSE levels as low as 0.1 per cent of supposed scrapie cases, for example—a level that would have significant consequences for public health—thousands of scrapie cases would need to be screened (see box).

Until the risk of BSE in sheep is better assessed, scientific advisers are left contemplating an unknown risk. If BSE had entered the sheep population in the 1980s it may still be present. SEAC considers that, whereas cattle are a dead-end host for BSE, there is evidence that in sheep BSE behaves like scrapie. Scrapie is transmitted from animal to animal and is endemic, and they suspect that BSE in sheep might behave similarly.

The BSE prion shares characteristics with scrapie in sheep. The distribution of BSE infectivity in experimentally infected sheep mirrors that of scrapie and not that of BSE in cattle, with infectivity being present in a wider range of organs. The possibility remains that BSE was introduced in the sheep flock in the mid-1980s and has persisted, says Almond.



Lambs to the slaughter: no cases of BSE have been found but how hard should we look?

PETE ADDIS

The wider distribution of infectivity of BSE in sheep also suggests that effective precautions to protect the public might be more difficult to make than in cattle, and that existing measures might be inadequate if BSE were to occur in sheep.

Slaughter

UK measures, based on SEAC recommendations, include the compulsory slaughter of sheep and goats with scrapie with compensation for farmers, and a ban on clinically affected animals — and the heads, spinal cords and spleens of all animals — entering the human or animal food chains. Similar measures have been adopted in France, but not throughout the European Union.

But SEAC acknowledges that, if BSE behaves like scrapie, it “would also be present in the lymph nodes, spleen and parts of the intestine in medium titres in animals in the preclinical phase”. “If you wanted to minimize risk you would have in practice to condemn the entire carcass,” says Almond.

Almond justifies SEAC's decision to stop short of such drastic recommendations. “To do nothing would be inappropriate,” he says, “while banning lamb would be ridiculous.” It would not be possible to justify killing the entire national flock when not a single case of BSE had been identified, he says. “We had to find a middle ground, which we call a ‘risk reduction strategy’ as opposed to a ‘risk minimization strategy.’”

If new research identified BSE in sheep, SEAC would consider stricter measures, according to one member of the committee, who defends what he describes as an “incremental approach” that adds new measures according to perceived need.

Lack of scientific data on the risks has placed SEAC in a position analogous to that of the scientists who advised the government at the beginning of the BSE crisis, admits one member. “We are back to square one,” and “in a very, very difficult position” in recommending appropriate courses of action.

Also in a difficult position is Britain's new Labour government. In its handling of the risk of BSE in sheep, it is in danger of repeating the mistakes made by its Conservative predecessors in their management of the BSE beef crisis, warns the UK Consumers' Association.

The association wrote to Tessa Jowell, the minister of state for public health, on 10 August complaining that the government was not keeping consumers properly informed about the BSE sheep issue, and excessively concentrating decision making in the hands of experts. The confidential letter contests SEAC's recent public recommendation that no further controls — such as a ban on lamb offals — are needed at present.

Sheila McKechnie, the head of the association, argues that SEAC's role should be limited to advising the Department of Health on

Urgent need for faster diagnostic test

Sheep BSE research is hampered by the lack of a highly specific diagnostic test suitable for field use.

Conventional strain typing, pioneered by Moira Bruce at the BBSRC's Institute for Animal Health in Edinburgh, discriminates prion strains well, but involves inoculating legions of mice with sheep extracts, and analysing incubation times and lesion profiles in the brains of mice that succumb. Experiments take up to two years, and each costs £30,000.

John Collinge, head of the Prion Disease Group at Imperial College School of Medicine in London, and a member of SEAC, thinks he has a faster and cheaper method, but complains that the UK agriculture ministry has so far failed to commit itself to the refinement and scaling-up of the technique needed if large numbers of sheep are to be tested.

The technique distinguishes prion strains on the basis of Western blot patterns of conformation and glycosylation of prion proteins, and Collinge says pilot studies demonstrate its feasibility in screening for BSE in sheep.

But many scientists, including Stan Prusiner, who received the Nobel prize last year for his work on prions, remain sceptical about the theoretical basis of the method. Others question its validity in sheep where interpretation of patterns is complicated by the large variety of prion strains and host prion genotypes. The electrophoretic profiles of the many sheep prion genotypes needed to establish a baseline for comparison has also not yet been established.

Such uncertainties have divided SEAC members over the use of the technique. “It is not coming out clean in a way that makes it easy for us on SEAC to make decisions,” says one SEAC scientist.

These arguments miss the point, says Collinge, who argues that, even if the theoretical basis for his technique is not clear, it is an empirical marker that works. He admits the technique needs refinement: “But that's MAFF's job not mine.”

Opinion on SEAC now seems to be swinging in Collinge's favour. Pattison says: “Glycotyping is likely to be one of the useful techniques in surveying the

strain characteristics of the agents recoverable from a large(ish) number of animals. I think we've reached the point where accumulation of further data is the way to inform ourselves about the usefulness of the technique rather than further theorizing.”

One compromise strategy that seems to be emerging from SEAC discussions is that, in the absence of a rapid specific test, a targeted search could use Collinge's technique to screen several thousand scrapie cases and identify suspicious signatures which could then be double-checked using the traditional strain-typing technique.

Some of the resistance to Collinge's technique seems, however, to be linked to fears that a preliminary diagnosis of BSE in sheep using this method may be interpreted by the public as firm evidence that BSE has passed to sheep. “The consequences of identifying a first case of BSE in sheep would be catastrophic so we need to be really sure [about the reliability of the identification],” says one member of the European Commission's independent Scientific Steering Committee. **D.B.**

the state of the science. Public health requires not only a scientific judgement but “a political and social one,” she argues, which is the responsibility of government and not experts. “I find this very distressing,” says McKechnie, “it is like history repeating itself”.

Advice

SEAC also needs to explain more clearly to the public the rationale of its advice to government, argues Harriet Kimbell, a lecturer at the Guildford College of Law, and the first consumer representative on SEAC. “The public should be able to make the same informed judgements on their eating habits as I, as a member of SEAC, can,” Kimbell claims SEAC is reluctant to make this cultural change — “it is not the most open of bodies,” she says. Pattison was unable to be reached for comment on this.

The letter to Jowell also called on the health department to consider advising parents not to feed lamb to young children, “who may not have been exposed to previous risks

of BSE and not have eaten lamb so far”. The letter argued that until the risk was better quantified this might be justified on the basis of “the public health principle of both precaution and proportionality”.

The letter added that the association rejected concerns that “to issue any kind of statement would result in media hysteria and cause major damage to the industry,” arguing that this was the policy taken by the previous government. SEAC dismissed the association's suggestion, and DOH has since said it does not intend to issue any further advice at this time.

The quality of the reasoning by some SEAC members is open to debate, however. One member criticizes the association's proposal, arguing that: “If you start saying don't feed lamb to children all hell will break loose.” Another says: “I did not understand why children would be at a higher risk than adults, there may be actual reasons to speculate on that, but those would hardly be sufficient to justify such selective precautions.” □

Science set for star role in Australian election campaign

[SYDNEY] Australia's opposition Labor party has taken the rare step for a political party of promoting support for research and development (R&D) as a key plank in its election campaign. National polls are expected sometime next month.

Labor's leader Kim Beazley has promised to restore a 150 per cent tax concession for industrial R&D if elected. The concession was cut to 125 per cent in 1996 by the present ruling coalition government of the Liberal and National parties.

Labor's proposal would be worth about A\$600 million (US\$330 million) over three years. The concession would be targeted at companies with fewer than 500 employees. Larger companies would qualify if they spent more than 2.5 per cent of turnover on R&D. Other firms would be eligible for a tax concession at the current rate.

Pay crisis deepens as Russian ruble plummets

[MOSCOW] Russia's financial crisis, the devaluing of its currency, and the sacking of the cabinet has left in tatters government promises to repay debts to scientists.

"The Ministry of Finance tells us that banks lack money and that it is impossible to pay salaries to scientists," says Vladimir Khlebodarov, chairman of the Russian Academy of Sciences' labour union. Khlebodarov says the union is planning to organize new, "more effective" actions later this month.

Meanwhile, the post of minister for scientific research remains vacant following the removal by President Boris Yeltsin of the cabinet headed by the prime minister, Sergei Kirienko.

British graduates turn their backs on teaching

[LONDON] The number of graduates applying to teach science in British schools has dropped sharply, according to figures released by the Graduate Teacher Training Registry. The National Association of Head Teachers describes the situation as "extremely serious".

Physics has suffered the biggest fall, with just 150 graduates accepting university teacher training places in physics, 39 per cent down on the 1997 figure. Graduates training to teach chemistry are down 21 per cent, with biology down 16 per cent and mathematics down 31 per cent.

The figures are a significant blow to a government campaign to attract more — and better qualified — graduates to teaching. A

discussion paper will be published later this year proposing the introduction of performance-related pay.

Deforestation blamed for Himalayan landslides

[NEW DELHI] Deforestation and unplanned housing may have caused last week's landslides in the western Himalayas that killed more than 300 trekkers and buried whole villages, according to scientists and environmentalists in India.

The Geological Survey of India has started mapping landslide-prone zones to help state authorities better target housing development. Scientists are also calling for a central facility to forecast landslides.

J. S. Samra, the director of the Central Soil and Water Conservation Research Institute in Roorkee, says India urgently needs a separate institution for research into mountain hazards.

Shaggy cloned dog story for Texas vets

[WASHINGTON] Researchers at Texas A&M University have been promised an anonymous donation of \$2.3 million to clone Missy, the donor's dog.

The research team — led by Mark Westhusin, associate professor of veterinary physiology and pharmacology at the university's College of Veterinary Medicine in College Station, Texas — has received an initial payment of \$500,000. They have already taken tissue biopsies of Missy, who was flown to Texas in May.

The project, announced last week by the university and the Bio Arts and Research Corporation of San Francisco, details its goals on its website (www.missyplexity.com). These include improving the "basic understanding of canine reproductive biology". The team also plans to lay the groundwork to clone "exceptional individual dogs of high societal value" after it has cloned Missy.

Germany 'should boost interdisciplinary research'

[MUNICH] Interdisciplinary research between universities, non-university institutes and industry in Germany should be strengthened in areas such as materials research, genetics and molecular medicine. This is the main recommendation of a report from a high level technology advisory group, the Technologierat.

The group, set up by Chancellor Helmut Kohl in 1995, comprises 17 experts from the academic community, industry and politics. Recommendations from its two previous reports — on information technology and biotechnology — have helped boost these

areas in Germany. The Technologierat also recommends that social sciences and humanities should increasingly contribute towards science and technology assessments.

The new report identifies science and technology as key to Germany's future economic development. It concludes that industrial exploitation of research results is unnecessarily delayed by the rigid and 'outmoded' separation of scientific disciplines and of basic and applied research.

In-utero gene therapy appears on NIH agenda

[WASHINGTON] The Recombinant DNA Advisory Committee (RAC) at the National Institutes of Health will discuss next month its first proposals for in-utero gene therapy.

The discussion has been prompted by two submissions by French Anderson, the gene therapy pioneer at the University of Southern California, who is preparing protocols to treat two diseases in utero: alpha-thalassaemia, and severe combined immunodeficiency caused by adenosine deaminase deficiency.

The RAC no longer has authority to approve protocols — this has been shifted to the Food and Drug Administration — but it still functions as a public forum for discussion of novel gene therapy protocols.

MIT to raze Rad Lab's home to the ground



[BOSTON] The Massachusetts Institute of Technology (MIT) last month received permission from the Cambridge Historical Commission to demolish its famous Building 20, where radar was perfected during the Second World War.

The dilapidated building was erected as a temporary structure in 1943 to house the Radiation Laboratory. In its time, the 'Rad Lab' was the biggest scientific collaboration in US history and a bigger effort than the Manhattan Project in terms of staff and funding. Building 20 was also the incubator of many other advances, including the first atomic clock. The site will provide a home for MIT's Department of Electrical Engineering and Computer Science, designed by architect Frank O. Gehry, who recently designed the Guggenheim Museum in Bilbao, Spain.

Too intelligent for our own good

Sir— Intelligence may be a deadly pathogen that, given sufficient time, will infect all habitable planets. It rapidly learns to obtain energy by reversing the geochemical fluxes that hitherto maintained the surface conditions of the planet stable and habitable.

Such changes may be very rapid. At the current incremental rate, forced by intelligent biota, atmospheric carbon dioxide will be 600 parts per million by the year 2050. This is a non-trivial increase over the 180 and 280 p.p.m. indicated by ice-core data for glacial and interglacial periods respectively over the previous 160,000 years¹.

In his Review Article in *Nature*, Timothy Lenton² did not discuss this most interesting aspect of Gaia — its current manifestation. I think it is now agreed that self-regulating feedback by the pre-sentient biosphere maintained, and may have produced, an atmospheric composition and a mean global surface temperature

conducive to the evolution of biota likely to develop intelligence. What is not clear is whether the sentient part of the biosphere will develop sufficient self-control to restore geochemical fluxes to the state necessary to maintain the surface of the Earth habitable.

In short, is intelligence a pathogen to which the biosphere can adjust, or is it terminal? Perhaps it is as well that we cannot yet observe other habitable planets, to discover what became of their sentient life!

Alan Longhurst

*Galerie l'Acadie, Place de l'Eglise,
46160 Cajarc, France*

Sir— Biology, geochemical cycles and climate are intimately linked. There are many feedbacks, both stabilizing and destabilizing, involving biology, geochemical cycles and climate. Obviously, systems with strong stabilizing feedbacks are more likely to persist than systems

lacking such stabilization. And species that are components of stable systems are more likely to persist than species that are components of unstable systems. These simple observations can explain the prevalence of important stabilizing feedbacks on Earth involving biology, geochemistry and climate.

Is there such a thing as a negative feedback involving biology, geochemistry and climate that is not “Gaian regulation”? If all such feedbacks are *ipso facto* Gaian, then “Gaian regulation” is but old wine in a new bottle.

Ken Caldeira

*Climate System Modeling Group,
Lawrence Livermore National Laboratory,
7000 East Ave, L-103, Livermore,
California 94550, USA
e-mail: kenc@llnl.gov*

1. Barnola, J. M., Raynaud, D., Korotkevich, Y. S. & Lorius, C. *Nature* **329**, 408–414 (1987).

2. Lenton, T. M. *Nature* **394**, 439–447 (1998).

Religious belief doesn't weaken scientific mind

Sir— Larson and Witham (*Nature* **394**, 313; 1998) address the controversial topic of religious belief. They quote Leuba, who attributed higher levels of disbelief and doubt to “superior knowledge, understanding, and experience”, and Atkins, who states “You can clearly be a scientist and have religious beliefs. But I don't think you can be a real scientist in the deepest sense of the word because they are such alien categories of knowledge”. Both opinions discredit religious scientists.

The essence of science, and moreover of the human mind, makes the process of acquiring knowledge infinite. Take matter as a simple example. The earliest attempts to define it divided matter into simple categories, then abstraction grew, and the concept of the atom was developed. We could not be content with indivisible entities, so we discovered protons, neutrons and electrons. We then began the search for subatomic particles. And so it continues, with no end in sight.

It is always tempting to ask new questions, which means that scientists are ultimately faced with a huge number of unanswered questions, and with the fact that each answer will probably create more doubts than it will solve. It is like floating in the middle of the Pacific Ocean: we can move a bit towards the edge, but we know that we will never reach it. This mysterious

immensity is what makes science so fascinating.

Once we realize that it is impossible to answer all the questions, there are two options: to accept this fact or to attribute the ‘answer’ to an almighty being — a God. Whichever option we choose does not make any difference to the quality of our research or to our fascination for life's mysteries. As long as we are still looking for the truth, we are on the right track.

Raul Camba

*Inorganic Chemistry Laboratory, University of
Oxford, South Parks Road, Oxford OX1 3QR, UK*

Make space for levity

Sir— Cosmologists are now getting data at rates greater than those conveniently measured in bits per millennium. *Nature* and *Science* have recently presented many interesting articles on the subject. But otherwise well written articles seem to have difficulty describing the effects of a greater than zero cosmological constant. Phrases such as “energy density of empty space”¹, “cosmic repulsion”², “large scale repulsive force”³, “vacuum energy”⁴, “mysterious repulsive force”⁵ and “cosmic antigravity”⁶ are inconsistent with each other, and less than clear.

I would like to propose that the effects of a greater than zero cosmological constant be referred to as ‘levity’. In the vernacular, levity is rather the opposite of gravity. It starts with ‘l’, the Latin equivalent of the

Hellenistic lambda. And, while I make no great claims of being a wordsmith, ‘levity’ or ‘levitational force’ seems to read better than any of the alternatives above.

In a field where there are terms such as ‘black hole’ and ‘big bang’, and where many researchers report feeling uneasy with a greater than zero cosmological constant, surely a little levity wouldn't hurt.

Dan Nagle

*12142 Purple Sage Court, Reston,
Virginia 20194-5621, USA
e-mail: dnagle@erols.com*

1. Battersby, S. *Nature* **392**, 548 (1998).

2. Coles, P. *Nature* **393**, 741 (1998).

3. Glanz, J. *Science* **280**, 1522 (1998).

4. Kamionkowski, M. *Science* **280**, 1397 (1998).

5. Glanz, J. *Science* **280**, 1008 (1998).

6. Glanz, J. *Science* **279**, 1298 (1998).

Citation heaven

Sir— I read the report in *Nature* with great excitement. At the 1999 *Nature* Conference on Citation Formats all scientific journals had agreed to use one of three citation formats from 1 January 2000. This was a transition agreement until 1 January 2010 when a single format would be adopted.

Sadly, I awoke from my dream to face another day of formatting my manuscript to the appropriate citation style, and daydreamed of how simple life could be.

Brian Dean

*Division of Molecular Schizophrenia, Mental Health
Research Institute, Parkville, Victoria, Australia*

Vocal gymnastics and the bird brain

Franz Goller

How do birds sing? The characteristic sound properties of birdsong were thought to be controlled by the central nervous system. But a new study shows that acoustic effects intrinsic to the avian vocal organ also contribute to zebra finch song.

For songbirds, singing is a complex behaviour that must be learned. It has stimulated rapidly advancing research in various disciplines, notably neurobiology and behavioural ecology¹, yet we still do not understand in detail how sound is produced by the birds' vocal organ, the syrinx. The main reason for this is that the syrinx is located at the base of the trachea (windpipe), making it relatively inaccessible to direct physiological studies² — the powerful, direct methods that have been successfully used to study sound production in the human voice box cannot easily be adapted to investigate the avian syrinx. So our ideas about sound generation in birds are based on indirect approaches, such as analysis of vocalizations and of the morphology of the syrinx, and theoretical models.

A report by Fee *et al.*³ on page 67 of this issue shows that a combination of indirect and direct approaches can help to overcome these difficulties. Their careful analysis of zebra finch (*Taeniopygia guttata*) song revealed linear and nonlinear phenomena, including switches from periodic to aperiodic or chaotic oscillations and period doubling (see box). Transitions from linear to nonlinear dynamics occurred rapidly (within 1 ms), without silent intervals between the two states, suggesting that the transitions arise from intrinsic properties of the vibrating components of the syrinx rather than from complex neural control.

Until now, it was assumed that the central nervous system directly controls the often intricate temporal pattern of song⁴. In birds, singing involves the expiratory muscles that line the body wall and generate pulses of increased air pressure by compressing the posterior air sacs (Fig. 1). These pulses define the coarse temporal pattern, which can be modified by activity of the syringeal muscles. These muscles are attached to the syrinx, and they turn sound production on and off by opening and closing the airways through the syrinx. The respiratory and syringeal muscles also control the acoustic structure of song, such as sound frequency and amplitude, and frequency modulation⁵. An elaborate network of brain areas controls the respiratory and syringeal muscles during song production⁶. But we now learn that intrinsic

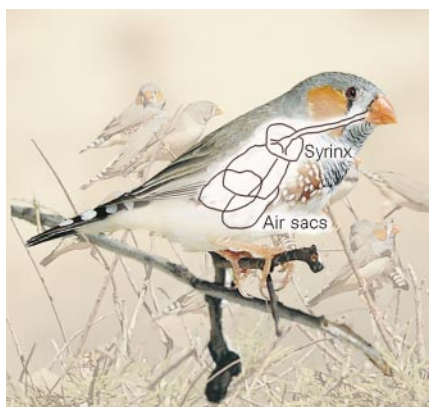


Figure 1 Vocal organs of the zebra finch, *Taeniopygia guttata*. Fee *et al.*³ have shown that birds don't rely on their brains as much as was thought to generate the complex acoustic patterns characteristic of birdsong. Instead, there is a nonlinear component that can be measured even when the bird vocal organ, the syrinx, is detached from the syringeal, respiratory and expiratory muscles.

mechanical properties of the syrinx can contribute to temporal and acoustic song patterns. These patterns are independent of complex central control, requiring a minimal contribution (in the form of slowly changing pressure) from the respiratory and vocal muscles.

Fee *et al.*³ discovered this by studying the vibratory behaviour of the zebra finch syrinx in an *in vitro* preparation. Sounds induced by drawing air through the excised syrinx

showed linear and nonlinear acoustic phenomena, similar to those observed in zebra finch song. Interestingly, transitions from linear to nonlinear dynamics occurred spontaneously as the pressure on the syrinx was slowly increased or decreased. The transitions probably depend on the elastic and oscillatory characteristics of the vibrating structures, the labia and medial tympaniform membranes. Using stroboscopic imaging, the authors showed that these linear and nonlinear oscillations came from the vibrating components of the syrinx. Because nonlinear effects can be generated in the isolated syrinx, they must be independent of complex control by the syringeal muscles. Furthermore, Fee *et al.* reproduced these nonlinear effects using a quantitative model of syringeal dynamics, providing additional theoretical support for their new hypothesis.

Young zebra finches learn to sing by copying the song of an adult male. Because the newly discovered transitions from linear to nonlinear dynamics are not generated by direct motor control, it would be interesting to know whether they are copied as accurately and frequently as other parts of the song. Fee *et al.* suggest that morphological differences between the syrinx in different individuals may induce transitions of the vibratory mode at different pressures. If so, then different birds account for these differences in their motor control, to generate the same acoustic control. But how does the central nervous system incorporate the intrinsic properties of the syrinx when a bird is learning to sing? Fee *et al.* propose that song learning requires a representation of syringeal dynamics. Such a level of sophistication may not be necessary, however, if birds can achieve the correct acoustic effect by learning pressure adjustment through trial and error.

These results should interest those studying the evolutionary and ecological aspects of vocal behaviour in birds. Because nonlinear phenomena add to the temporal and acoustic complexity of song, we need to ask whether these contributions have any perceptual significance — complexity of vocal signals may convey information about the

The dynamics of song

If an effect is described as linear, the amplitude of the response is directly proportional to the amplitude of the stimulus. So in a vocal organ, for example, if the oscillating valve followed linear dynamics, a two-fold increase in air pressure would result in doubled acoustic flow. Nonlinearity, on the other

hand, means that there is no such direct proportionality, resulting in a more complex relationship between pressure and flow.

In a periodic oscillation, the period — that is, the duration from peak to peak of the sound wave — repeats regularly. But aperiodic (or chaotic) oscillations

are characterized by irregularity, and there are no repeating periods at all in the most extreme case. Period doubling is a characteristic of nonlinear dynamical systems. It is characterized by a change in the frequency of the oscillations, such that the spacing between spectral components is halved.

F.G.

sender. In some species of bird, for example, acoustic versatility of song is an indicator of male reproductive fitness⁷. So, this may be important for choice of mate and encounters between members of the same sex. If peripherally generated acoustic structure requires a less precise motor control than complex sound modulation controlled by the action of muscles, is it also weighed differently by a listener who is trying to work out the 'quality' of the singer?

The findings are also of practical importance for researchers trying to quantify the quality of birdsong. Our assessment of song complexity is tightly linked to our knowledge of sound-producing mechanisms, and now that peripheral contributions to song structure must be added to the picture, the task has become even more challenging.

Finally, there remains the question of whether nonlinear dynamics might also be involved in singing by other species of bird. I suspect that, as the news spreads, more examples of nonlinear effects contributing

to the temporal and acoustic pattern in bird vocalizations will be described. Nonlinearity is also well recognized in the physiology of the human vocal organ, albeit often in connection with voice disorders⁸. But to those who suffer from a roughness of voice, it must be of little comfort to know that nonlinearity can also be a mechanism to enhance vocal properties. □

Franz Goller is in the Department of Biology, University of Utah, 201 South Biology, Salt Lake City, Utah 84112, USA.

1. Konishi, M., Emlen, S., Ricklefs, R. E. & Wingfield, J. C. *Science* **246**, 465–472 (1989).
2. Gaunt, A. S. & Nowicki, S. in *Animal Acoustic Communication* (eds Hopp, S. L., Owren, M. J. & Evans, C. S.) 291–321 (Springer, Berlin, 1998).
3. Fee, M. S., Shraiman, B., Pesaran, B. & Mitra, P. P. *Nature* **395**, 67–71 (1998).
4. Konishi, M. *Curr. Opin. Neurobiol.* **4**, 827–831 (1994).
5. Goller, F. & Suthers, R. A. *J. Neurophysiol.* **75**, 867–876 (1996); **76**, 287–300 (1996).
6. Wild, J. M. *J. Neurobiol.* **33**, 653–670 (1997).
7. Hasselquist, D., Bensch, S. & von Schantz, T. *Nature* **381**, 229–232 (1996).
8. Herzel, H. in *Vocal Fold Physiology* (eds Davis, P. J. & Fletcher, N. H.) 63–75 (Singular Publishing, San Diego, 1996).

Quantum mechanics

Where the weirdness comes from

Peter Knight

More than 60 years after the famous debate between Niels Bohr and Albert Einstein on the nature of quantum reality, a question central to their debate — the nature of quantum interference — has resurfaced. Dürr, Nonn and Rempe, reporting on page 33 of this issue¹, have used an atom interferometer to show that Schrödinger's concept of 'entanglement' between the states of particles is the key to wave-particle duality, and to under-

standing much that is weird about quantum mechanics. This is quite different from the usual textbook explanation of duality in terms of unavoidable 'measurement disturbances'. It confirms that entanglement is essential in establishing quantum weirdness and in the emergence of classical behaviour at larger scales.

Quantum entities can behave like particles or waves, depending on how they are observed. They can be diffracted and pro-

duce interference patterns (wave behaviour) when they are allowed to take different paths from some source to a detector — in the usual example, electrons or photons go through two slits and form an interference pattern on the screen behind. On the other hand, with an appropriate detector put along one of the paths (at a slit, say), the quantum entities can be detected at a particular place and time, as if they are point-like particles. But any attempt to determine which path is taken by a quantum object destroys the interference pattern. Richard Feynman described this as the central mystery of quantum physics.

Bohr called this vague principle 'complementarity', and explained it in terms of the uncertainty principle, put forward by Werner Heisenberg, his postdoc at the time. In an attempt to persuade Einstein that wave-particle duality is an essential part of quantum mechanics, Bohr constructed models of quantum measurements that showed the futility of trying to determine which path was taken by a quantum object in an interference experiment. As soon as enough information is acquired for this determination, the quantum interferences must vanish, said Bohr, because any act of observing will impart uncontrollable momentum kicks to the quantum object. This is quantified by Heisenberg's uncertainty principle, which relates uncertainty in positional information to uncertainty in momentum — when the position of an entity is constrained, the momentum must be randomized to a certain degree.

This explanation in terms of the uncertainty principle has become a talisman for some, but it has left others uneasy, as it views the measurement and momentum kicks as 'locally realistic' — in other words, as idealized classical measurements, rather than quantum mechanical phenomena themselves. This is a dangerous position, and it has led to debate in this journal between a group centred on the Max-Planck Institute for Quantum Optics² and one in Auckland³, on whether momentum kicks are necessary to explain the two-slit experiment. Obviously, momentum is involved, because a diffraction pattern is a map of the momentum distribution in the experiment. But *how* is it involved? Is it everything, as Bohr would have claimed?

This is the question addressed by Dürr *et al.*¹, who have studied the interference fringes produced when a beam of cold atoms is diffracted by standing waves of light. Their interferometer displays fringes of high contrast — but when they encode within the atoms information as to which path is taken, the fringes disappear entirely. The internal labelling of paths does not even need to be read out to destroy the interferences: all you need is the option of being able to read it out.

The key to this new experiment is that

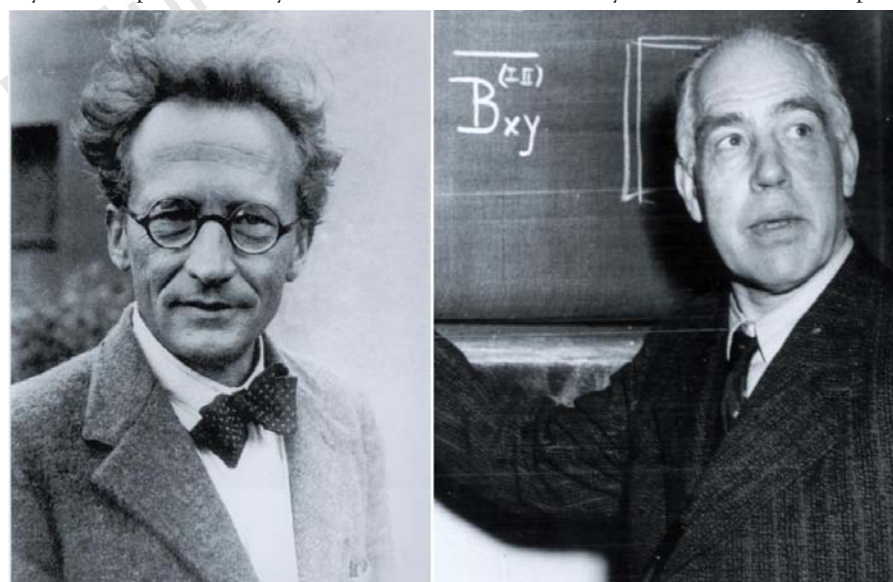


Figure 1 Erwin Schrödinger (left) and Niels Bohr. Bohr claimed that a momentum kick, imparted by any measurement of particle position, could explain the disappearance of quantum interference in 'two-slit' experiments. A new experiment¹ shows that this effect is too small, and the disappearance must instead be explained using Schrödinger's 'entanglement' between quantum states.

although the interferences are destroyed, the initially imposed atomic momentum distribution leaves an envelope pattern (in which the fringes used to reside) at the detector. A careful analysis of this pattern finds that it has not been measurably distorted by a momentum kick of the type invoked by Bohr, and therefore that any locally realistic momentum kicks imparted by the 'which-path' measurement are too small to be responsible for destroying interference.

So some other phenomenon is at work here. It is 'entanglement', a peculiar but basic feature of quantum mechanics introduced by Erwin Schrödinger in 1935. Individual quantum-mechanical entities need have no well-defined state; they may instead be involved in collective, correlated ('entangled') states with other entities, where only the entire superposition carries information. That may apply to a set of particles, or to two or more properties of a single particle.

In the new experiment¹, the directional information is encoded by manipulating the internal state of an atom with a microwave field, which entangles the atom's momentum with its internal electronic state. Like all such entangled states, the constituent parts lose their separate identity. But the attachment of a distinguishable electronic label to

each path means that the total electronic-plus-path wavefunction along one path becomes orthogonal to that along the other, and so the paths can't interfere. Entanglement is the source of the weirdness.

We already know that quantum interference is fragile: we do not see everyday objects exhibiting wave-particle duality. In the past 15 years or so we have learned that the disappearance of quantum weirdness has to do with the entanglement of quantum objects with their environment, an environment so complex and so large that correlations are set up that we are unable to track. We lose information, and thus the coherence responsible for interference. The larger the object, the faster this loss of coherence⁴.

In the new experiment, entanglement leads to a way of distinguishing paths and destroying interferences. But a great deal of work must still be done to quantify this. What degree of information on path distinguishability produces what degree of fringe diminishment? □

Peter Knight is at the Blackett Laboratory, Imperial College, London SW7 2BZ, UK.
e-mail: peter.knight@ic.ac.uk

1. Dürr, S., Nonn, T. & Rempe, G. *Nature* **395**, 33–37 (1998).
2. Scully, M. O. *et al.* *Nature* **351**, 111–116 (1991).
3. Storey, B. E. *et al.* *Nature* **367**, 626–628 (1994).
4. Haroche, S. *Physics Today* July 36–42 (1998).

Cancer

Has the smart bomb been defused?

Steven P. Linke

A goal of cancer research is to develop therapies that can selectively kill tumour cells without adversely affecting normal cells — this is crucial for both the short-term comfort and long-term survival of patients. One such agent is a genetically engineered adenovirus from ONYX Pharmaceuticals, called ONYX-015, which is thought to selectively replicate in

(and kill) tumour cells deficient for the p53 tumour-suppressor pathway, having little toxicity to normal cells^{1,2}. Because p53 is mutated in over half of all human cancers, and is probably functionally deficient in many others owing to mutations in associated genes, ONYX-015 could prove useful in treating a range of tumours. But a report by Hall *et al.*³ in *Nature Medicine* raises a serious

question about the efficacy of this virus.

In stark contrast to initial results published by ONYX^{1,2}, Hall and colleagues suggest that adenoviruses *require* p53 to kill cells efficiently. They propose that ONYX-015 would selectively destroy normal cells. Yet the claims by both groups about the dependence on p53 are subject to alternative interpretations, so no definitive conclusions can be reached at present.

When activated by stresses such as DNA damage, p53 can induce the transcription of genes that mediate anti-proliferative responses, including temporary or permanent growth arrest, or programmed cell death (apoptosis)⁴. Another tumour-suppressor protein, retinoblastoma (Rb), can work in conjunction with p53 to inhibit progression to the replicative (S) phase of the cell cycle. Infection with DNA tumour viruses can also induce anti-proliferative responses, so viruses encode oncoproteins that inactivate both tumour suppressors. The adenovirus early 1A (E1A) protein inactivates Rb, forcing cells into S phase and allowing the virus to replicate. This inappropriate entry into the cell cycle would normally activate p53, causing growth arrest or apoptosis before the virus could replicate. But two other adenoviral proteins, E1B 55K and E1B 19K, help to inactivate the p53 pathway: E1B 55K binds to and inhibits transcriptional activation by p53, whereas E1B 19K inhibits p53-dependent apoptosis⁵ (Fig. 1).

The ONYX group deleted the gene that encodes E1B 55K, to produce the ONYX-015 virus. Theoretically, normal cells (with an intact p53 pathway) would be able to mount an anti-proliferative response after infection with ONYX-015, owing to the absence of E1B 55K. Because E1B 19K would still be present to inhibit p53-dependent apoptosis, cells would presumably undergo p53-dependent growth arrest instead, maintaining their viability. But p53-deficient tumour cells would be unable to mount such a response. Instead, they would cycle and support replication of the virus. Moreover, unlike other gene-therapy vectors, ONYX-015 can replicate in host cells, so it can spread from the site of delivery to surrounding tumour cells. These theories indicate that ONYX-015 is a potential 'smart bomb' against cancer⁶. Indeed, the ONYX group found^{1,2} that ONYX-015 replicates in tumour cell lines with mutant or no p53, but not in primary, non-immortalized human strains of endothelial, epithelial or fibroblast origin. Another group has extended these results to a few other cell types⁷.

In contrast, Hall *et al.*³ and others⁸ now suggest that the ability of wild-type and E1B-deleted adenoviruses to replicate does not correlate closely with the p53 status of the cells. Whereas the ONYX studies equate viral replication with cytolysis (cell destruction), Hall and colleagues distinguish these pro-

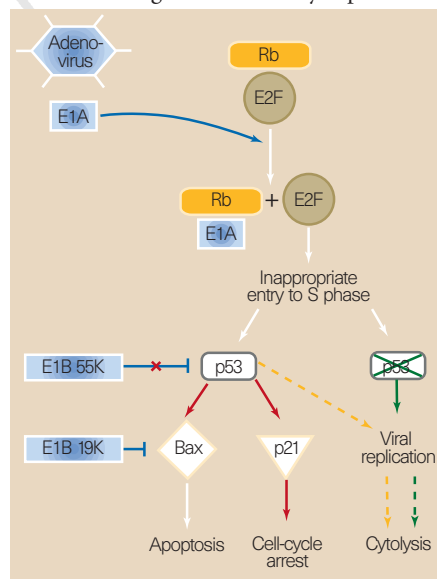


Figure 1 Influence of adenoviral E1A/E1B proteins (blue) and cellular p53 status on viral replication and cell death. E1A binds to retinoblastoma (Rb), releasing the cellular transcription factor E2F and promoting inappropriate entry into S phase. E1B prevents p53-dependent apoptosis (mediated by Bax, for example) or growth arrest (mediated by p21, for instance) in response to this stress. According to the ONYX group^{1,2}, p53-proficient cells would undergo growth arrest (red arrows) after infection with ONYX-015, whereas p53-deficient cells would continue to cycle, supporting viral replication and eventual cytolysis (green arrows). Hall *et al.*³ suggest that viral replication is supported in cells, regardless of their p53 status, whereas cytolysis requires wild-type p53 (yellow arrows). Dashed lines denote points of disagreement between the models.

cesses. They propose that although the virus can replicate in all cells, regardless of p53 status, wild-type p53 is required for cytolysis. They showed cell death in all seven of the tumour lines studied with wild-type p53, but in only three out of nine p53-deficient tumour lines tested.

These studies seem to come to opposite conclusions. However, a common weakness is the reliance on immortalized tumour cell lines of different genetic backgrounds and tissue types. As well as the changes in p53, such cell lines contain many dissimilar genetic alterations, and this may explain why many of the results from both groups contradict their own models. For example, ONYX showed that their mutant virus replicates in some normal cells, and kills many tumour lines with wild-type p53. Although they establish that the mutant virus replicates in and kills tumour cell lines more efficiently than primary cell strains^{1,2,7}, the data do not support a dependence on p53.

Another explanation for the disparate results is the different end points that the two groups used to determine cell death. Whereas the ONYX group assayed nine days after viral infection, Hall and co-workers assayed after three days. In fact, these authors acknowledge that many lines with mutant p53 are dead at ten days, but they attribute this to non-physiological 'metabolic shut-down'. To resolve these conflicting data, experiments need to be done, with a consistent end point, in normal cells where the only alteration is in p53 status and/or in tumour cell lines that were null for p53, but in which p53 has been restored.

What characteristics, other than p53 status, might explain the fact that the mutant virus replicates in, and lyses, many tumour lines much more efficiently than normal cells? Goodrum and Ornelles⁸ have used synchronized cultures of HeLa cells to show that

the replication and cell-killing efficiency of an E1B-deficient adenovirus similar to ONYX-015 is correlated with the percentage of cells in S phase of the cell cycle at the time of infection. Tumour lines generally have more S-phase cells than do primary strains, so they might support stronger viral replication and cytolysis — as has been observed. Moreover, in clinical trials, ONYX-015 replicated much less in human tumours than in mouse xenografts, perhaps because the S-phase fractions are lower in the human tumours.

With the available data, only rough correlations can be made between p53 status and the effectiveness of ONYX-015 in eradicating tumour cells. However, if current trials are successful, this point may become moot from a clinical standpoint. Early data indicate that ONYX-015 can shrink tumours with few effects on normal tissue, even at high doses. But the role of the E1B deletion and the long-term effects on normal tissues are open to question. Thus, Hall *et al.* have not defused the 'smart bomb', but their study raises questions about adenovirus biology that warrant caution in the use of replication-competent adenoviruses as cancer treatments. □

Steven P. Linke is at the Laboratory of Human Carcinogenesis, National Cancer Institute, National Institutes of Health, 9,000 Rockville Pike, Bethesda, Maryland 20892-4255, USA.
e-mail: slinke@helix.nih.gov

1. Bischoff, J. R. *et al.* *Science* **274**, 373–376 (1996).
2. Heise, C. *et al.* *Nature Med.* **3**, 639–645 (1997).
3. Hall, A. R., Dix, B. R., O'Carroll, S. J. & Braithwaite, A. W. *Nature Med.* **4**, 1068–1072 (1998).
4. Wahl, G. M., Linke, S. P., Paulson, T. G. & Huang, L.-c. in *Cancer Surveys Vol. 29. Checkpoint Controls and Cancers* (ed. Kastan, M. B.) 183–219 (Cold Spring Harbor Lab. Press, Plainview, NY, 1997).
5. Yeh, P. & Perricaudet, M. *FASEB J.* **11**, 615–623 (1997).
6. Lowe, S. W. *Nature Med.* **3**, 606–608 (1997).
7. Freytag, S. O., Rogulski, K. R., Paielli, D. L., Gilbert, J. D. & Kim, J. H. *Hum. Gene Ther.* **9**, 1323–1333 (1998).
8. Goodrum, F. D. & Ornelles, D. A. *J. Virol.* **71**, 548–561 (1997).

as we are to the Sun. So given an unexplained source of radiation, we should first look to the Sun, or failing that, the Galaxy. If these are not effective radiators in the waveband concerned, then the distant Universe — typically thousands of megaparsecs (Mpc) from us — should dominate.

The magnetic fields of the Sun's atmosphere rarely accelerate particles to more than a few hundred million electron volts. Astrophysicists believe that most cosmic rays are accelerated up to their high energies by the magnetic fields and shockwaves of supernova remnants in our galaxy. But this mechanism has its limits too, and we must look elsewhere for the production of cosmic rays above about 10^{18} eV. Could they be coming from the vast plasma structures of radio galaxies and quasars?

Above energies of about 5×10^{19} eV, cosmic-ray particles lose energy continually by interactions with the 2.7 K cosmic microwave background radiation^{2,3}, and few cosmic rays above 10^{20} eV should be able to reach us from sources more than about 50 Mpc away. Protons and atomic nuclei — even, by a different process, photons — should all suffer attenuation. If we are looking at cosmic rays from distant radio galaxies or quasars, we should see the spectrum cut off at this energy.

But we don't. The Haverah Park experiment in Yorkshire⁴ has seen no evidence of a turn down near 10^{20} eV. Worse, in 1995 the air-fluorescence 'Fly's Eye'⁵ detector reported seeing a single particle of 3×10^{20} eV — by far the most energetic particle ever seen. In all cases, the statistical samples were small, but now, with nearly seven years of data¹ from the largest cosmic ray detector — the 100 km² Akeno giant air shower array (AGASA) in Japan — there is still no fall-off seen. So most of the cosmic rays above 10^{20} eV must have travelled less than 50 Mpc.

Could there by chance be an extraordinarily strong extragalactic source within this short distance, that outshines all else? The arrival directions of the observed cosmic rays above 4×10^{19} eV (Fig. 1, overleaf) militate against that idea — they are distributed widely over the sky, and there is no significant preference for the concentrations of galaxies^{6,7} within 50 Mpc. That even allows for the presence of magnetic fields in intergalactic space, which would bend the paths of these cosmic rays by 2 to 8 degrees, according to our best guess. Nevertheless, one cannot yet rule out the possibility that there is a much stronger magnetic field in some region which could tangle up the particle paths, camouflaging a unique source only a few Mpc away — just possibly the radio galaxy Centaurus A.

The simplest interpretation at present, though, is that these highest-energy cosmic rays come from our Galaxy, from sources in a large Galactic halo (as their arrival directions

High-energy astrophysics

Cosmic rays without end

Michael Hillas

Cosmic rays are highly energetic particles — mostly protons and atomic nuclei — that strike the Earth from space. They were first detected in 1912, but an understanding of their origin has been slow to arrive. In 1946, for example, Georges Lemaître proposed that they had all come from the super-radioactive decay of a single 'primeval atom' from which the Universe started — a forerunner of the idea of the Big Bang — and that they should therefore provide unique evidence of these beginnings. Nowadays, it seems probable that most cosmic rays are produced in the remnants of supernova explosions in our galaxy. But the most energetic cosmic rays of all, with ener-

gies above $\sim 10^{19}$ electron volts, are different: they appear to be coming from far outside our galactic disk and are very hard to account for. Now a report by Takeda *et al.*¹ in *Physical Review Letters* seems to have settled a controversy about the number of cosmic rays at the highest energies. This sharpens the problem of finding a source for the particles, and makes it appear quite possible that Lemaître wasn't far wrong.

In a city street, lit by a row of lamps, we find it natural that most of the illumination comes from the nearest lamp. But in a three-dimensional array of lamps like the Universe, distant sources produce most of the light, unless we are very abnormally close to one —

NATURE

100 YEARS AGO

The little Warwickshire village of Stockton, ploughed and excavated by three manufacturing cement firms, has long yielded to collectors choice specimens of Lower Middle Lias fossils. Its late rector educated the quarrymen by lectures and in conversation [and] left them with a prediction that a perfect monster would some day be unearthed... A week or two ago, the prediction was fulfilled, and the advice remembered. The wielder of a pickaxe suddenly announced that he was "grappling along a lot of backbones"; the work was stopped, the foreman summoned, and slowly with due precaution a noble Ichthyosaurus was uncovered. He lies 45 feet below the surface; 20 feet in length and 3 feet 10 inches across.



From *Nature* 1 September 1898.

50 YEARS AGO

From September 13 to September 17 the American Association for the Advancement of Science will hold in Washington, D.C., the centennial celebration of its founding... Beginning with 1900 each member received with his membership a subscription for the weekly journal *Science*, and since 1915 a subscription for either *Science* or *The Scientific Monthly*. At the beginning of this calendar year (1948) the annual membership dues, including a subscription for either of the two journals, became 6.50 dollars a year. From *Nature* 4 September 1948.

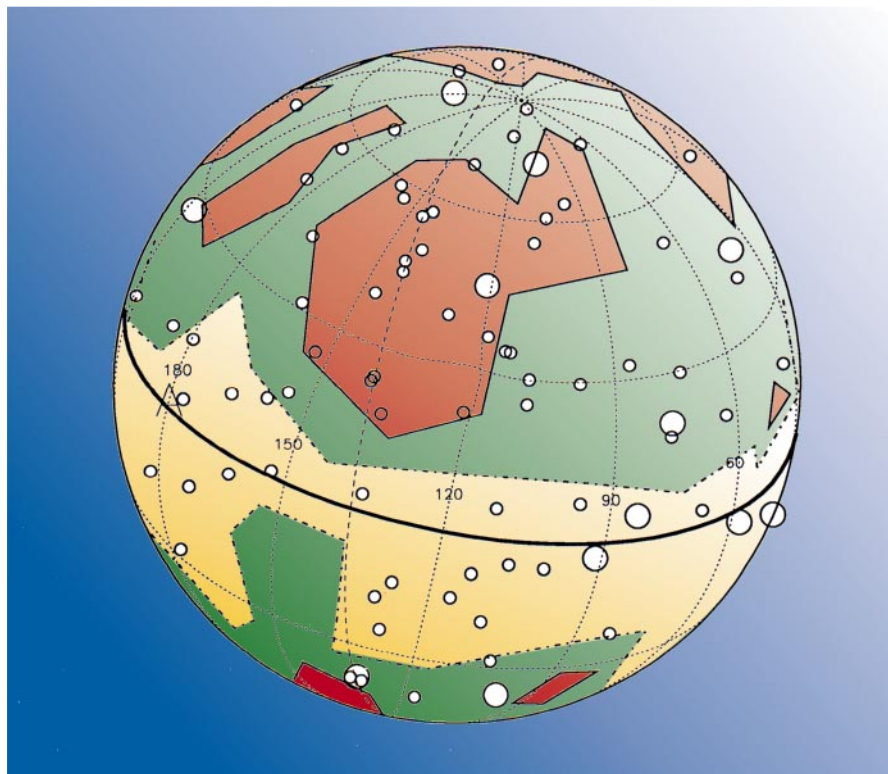


Figure 1 The high-energy cosmic-ray sky. Circles show the arrival directions of cosmic-ray particles; those above 10^{20} eV are larger. There are concentrations of galaxies in the nearby Universe (red directions) and voids (yellow); if the cosmic rays were coming from radio galaxies or quasars we would expect some bias towards these directions. Similarly, we would expect a bias towards the Galactic plane (thick black line) if they came from objects in our Galaxy's disk. Neither bias is seen, and the simplest solution is that the highest-energy cosmic rays come from sources in our Galactic halo — perhaps from the decay of exotic supermassive particles. (Adapted from Uchihori *et al.*⁶)

show no imprint of the Galactic disk). The Galactic halo itself is probably not a site of enough magnetic activity to generate such energetic protons. But most astronomers believe it to be filled with dark matter. Berezhinsky⁸ and Birkel and Sarkar⁹ have proposed two ways in which super-heavy particles might form in the very early Universe (echoes of Lemaître), and cluster in galactic haloes, eventually decaying with a half-life greater than the age of the Universe and generating cascades of high-energy protons, gamma rays and neutrinos. These speculations are not yet closely constrained by observations. But, if the particles have a mass about 10^{12} times the mass of a proton, the resulting decay spectrum is very reasonable⁹.

The paper by Takeda *et al.*¹ redresses another curious imbalance: there are now more particles that have been observed above 5×10^{19} eV than theoretical papers attempting to explain them. Many attempts have been made to bring particle physics and the early Universe to bear on the problem⁸, including the possibility of exotic particles that can pass unscathed through the primeval radiation, but most fail to match the known facts.

How fortunate, then, that the interna-

tional Auger project is planning to set up its first detector in Argentina. The planned 3,000-km² array of particle detectors will map out the spectrum and arrival directions of cosmic rays above 10^{19} eV from the southern sky. It should show whether the source distribution has a maximum — in the direction of the Galactic halo's centre or Centaurus A or elsewhere — and whether many gamma rays and neutrinos are mixed into the cosmic ray flux at these extreme energies, which would be evidence of massive-particle decays. □

Michael Hillas is in the Department of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK.

e-mail: a.m.hillas@leeds.ac.uk

1. Takeda, M. *et al.* *Phys. Rev. Lett.* **81**, 1163–1166 (1998).
2. Greisen, K. *Phys. Rev. Lett.* **16**, 748–750 (1966).
3. Zatsepin, G. T. & Kuz'min, V. A. *JETP Lett.* **4**, 78–80 (1966).
4. Lawrence, M. A., Reid, R. J. O. & Watson, A. A. *J. Phys. G: Nucl. Part. Phys.* **17**, 733–757 (1991).
5. Bird, D. J. *et al.* *Astrophys. J.* **441**, 144–150 (1995).
6. Uchihori, Y. *et al.* in *Extremely High Energy Cosmic Rays: Astrophysics and Future Observatories* (ed. Nagano, M.) 50–60 (Univ. Tokyo, 1996).
7. Medina Tanco, G. *25th Int. Cosmic Ray Conf., Durban* (eds Potgeiter, M. S., Raubenheimer, B. C. & van der Walt, D. J.) **4**, 477–480 (Potchefstroomse Univ., 1997).
8. Berezhinsky, V. preprint hep-ph/9802351 on xxx.lanl.gov.
9. Birkel, M. & Sarkar, S. preprint hep-ph/9804285 on xxx.lanl.gov.

Evolutionary biology

Debatable homologies

Diethard Tautz

“Homology: Similarity in structure of an organ or a molecule, reflecting a common evolutionary origin....” This is the definition of homology in a contemporary textbook¹. But although the concept lies at the heart of much of biology, it has become increasingly elusive. Has it therefore become a word ripe for burning, as J. Maynard Smith (Univ. Sussex) remarked at a meeting* on the topic? Or is it simply enough to know that homology exists — even though we cannot define it — as D. Wake (Univ. California, Berkeley) suggested in 1994 in a review² of a book³ commemorating the 150th anniversary of Richard Owen’s introduction of the term?

Owen’s original concept of homology did not have a phylogenetic (evolutionary) perspective. It was based on the philosophy of identifying and grouping similarities in nature, the ‘scala naturae’. Such thinking was exemplified by Charles Bonnet, who, in 1764, derived the following similarity series: fish — flying fish — aquatic birds — birds — bats — flying squirrels — tetrapods — monkey — man (A. Panchen, Univ. Newcastle upon Tyne). It is one of the great achievements of applying the principle of homology that we can now confidently reject the idea that this series reflects phylogenetic succession. On the other hand, however, we still do not know the true answer on the phylogenetic descent of tetrapods, as homology concepts tend to fail when it comes to tracing evolutionary novelties.

Further challenges arise from molecular comparisons of developmental genes. The finding that highly conserved ‘master regulator’ genes such as *Pax6/eyeless* are involved in eye development in diverse phyla⁴ brings into question the idea of the independent evolution of eyes. Similarly, the expression of *distal-less* homologues in insect and vertebrate legs, as well as other body outgrowths⁵, raises the issue of how often appendages can arise independently. The problem goes even further — with biotechnology companies proposing to use such organisms as *Drosophila* or zebrafish as models for understanding human diseases, it will eventually be investors and shareholders who have to be convinced that invoking homology is valid.

Taking phylogenetic continuity alone as the prime criterion for homology causes the problem of circularity, because phylogenies are themselves deduced from homologous characters. This of course can be avoided by using a different set of characters for the phylogeny, such as those derived from DNA

sequences. But even in well-established phylogenies, it often emerges that independent evolution of similar characters (convergence) must have often occurred. There could, therefore, be dormant genes or ‘latent homologies’ that are not expressed in a particular ‘stem’ species, but that have been regained after further speciation and give the erroneous impression of convergence (A. Meyer, Univ. Konstanz).

Strict correspondence with phylogeny might also be less important if homologues are viewed as coherent units of phenotypic evolution with constraints on the developmental and variational properties (G. Wagner, Yale Univ.). Applying such criteria, Wagner proposed a beautiful solution to the old question of the way in which the digits on the limbs of the urodeles, the newts and salamanders, are homologous to those of the other tetrapods. In tetrapods, digits 3 and 4 are the first to develop and are also the ones that are most stable against perturbations. In the urodeles, the same seems to be the case for digits 1 and 2. So if 3 and 4 are homologous to 1 and 2, digit 3 would be equivalent to digit 5 in tetrapods and digits 4 and 5 in the urodeles would be novel structures, albeit serially homologous to the other digits. Closer study of developmental characteristics and comparative analysis of *Hox* gene expression seem to confirm this view.

In molecular studies, the question of phylogenetic continuity becomes even more blurred. As genes can duplicate in the genome, it has long been clear that orthologous genes (which are related by phylogenetic common descent) have to be distinguished from paralogous genes that result from duplications within a genome. But if gene loss and re-duplications have to be taken into

account, things become even more complex and new terminology may be required (P. Holland, Univ. Reading). Thus, sequence similarities alone are not sufficient to infer homology of structures and developmental processes. Instead, one should look for the conservation of whole regulatory networks (E. Abouheif, State Univ. New York), a concept that comes close to that of viewing homologues as coherent units.

But can we be sure that homologous morphological structures are made by homologous gene networks? For example, closely related species of sea urchin can develop directly into adults, or go through a feeding larva first, yet end up with the same adult body plan. These species show different embryonic cell lineages and differences in patterns of gene expression during early development (R. Raff, Indiana Univ.). Intriguingly, however, the embryos of hybrids between such species develop partly in a composite and partly in radically new ways. Further study of these hybrids should help in understanding evolutionary dissociations of genotype and phenotype.

It has often been hypothesized that changes in regulatory interactions are involved in dissociations of this sort, but practical work on the evolution of regulatory modules (enhancers) has been scarce. Again, research on sea urchins might pave the way forward. Their enhancers seem to consist of sub-elements, which can be combined and integrated in different ways⁶, possibly making them perfect targets of evolutionary change (G. Wray, State Univ. New York). But could this mean that we eventually have to identify homologues in the individual parts of complex enhancers to understand the homology of morphological structures? This would indeed make the word ripe for burning.

Homology, then, is an idealized principle that works under idealized conditions, but such conditions almost never apply. This is reminiscent of the concept of the

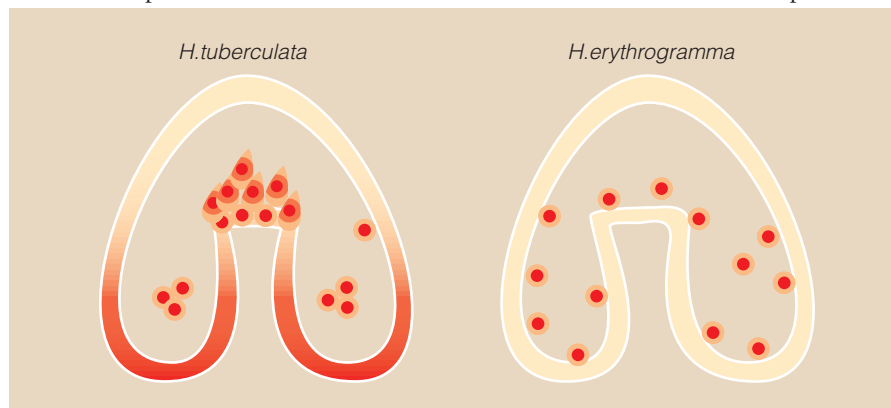


Figure 1 Drastic changes of cytoplasmic actin-gene expression (red) in two closely related sea-urchin species. *Helicidaris tuberculata* and *H. erythrogramma* split only about 10 million years ago, but they show different modes of development and regulation of actin genes, even though both develop into the same juvenile stage. Analysing hybrids between such species may help us to understand the evolution of gene networks in early development. (Adapted from ref. 7.)

* Homology, Novartis Foundation, London, 21–23 July 1998. Proceedings to be published by Wiley.

Hardy–Weinberg equilibrium in population genetics. It exists only under idealized conditions, but it is the tracing of the reasons for deviation from these conditions that is central to understanding the evolution of populations. Could one apply a similar thinking to homology? □

Diethard Tautz is in the Zoologisches Institut der Universität München, Luisenstrasse 14, 80333 München, Germany.
e-mail: tautz@zi.biologie.uni-muenchen.de

1. Alberts, B. et al. *Molecular Biology of the Cell*, 3rd edn (Garland, New York, 1994).
2. Wake, D. B. *Science* **265**, 268–269 (1994).
3. Hall, B. K. (ed.) *Homology — The Hierarchical Basis of Comparative Biology* (Academic, San Diego, California, 1994).
4. Quiring, R., Walldorf, U., Kloter, U. & Gehring, W. J. *Science* **265**, 785–789 (1994).
5. Panganiban, G. et al. *Proc. Natl Acad. Sci. USA* **94**, 5162–5166 (1997).
6. Yuh, C. H., Bolouri, H. & Davidson, E. H. *Science* **279**, 1896–1902 (1998).
7. Kissinger, J. C. & Raff, R. A. *Dev. Genes Evol.* **208**, 82–93 (1998).

Celestial mechanics

Orbits of all sorts

Donald G. Saari

Can we predict how three bodies, attracting each other by gravity, should move? Isaac Newton is said to have got a headache trying to solve the three-body problem — and in the form of the more general n -body problem, it has been giving mathematicians a headache ever since. Two centuries after Newton, Henri Poincaré exposed the source of the difficulty^{1,2}. Once there are three bodies or more, the system is chaotic: the motion depends sensitively on initial conditions, and is highly unpredictable. But that is not a dead end, for one can still make progress by concentrating on special solutions. On page 51 of this issue³, Gregory Buck introduces a new set of approximate special solutions, in which a very large number of bodies progress around a looped path. To appreciate the enormity of the new class of solutions, imagine any looped, tangled or knotted arrangement of wires. As long as the wires have no loose ends,

this is one of Buck's n -body orbits.

One way to find special n -body solutions is to look for 'central configurations'. Intuition rightly tells us that, for simple solutions to occur, the bodies must have a special symmetrical configuration where competing forces conveniently cancel, so that each body experiences an inverse-square attraction towards some central point, like that in a two-body system. In other words, as far as each body is concerned, the remaining bodies could be clumped together at the centre of mass of the whole system. That condition imposes strict constraints on the allowable masses and configurations, but once such a 'central configuration' is found, two-body solutions can be mimicked by selecting initial velocities so that the configuration does not change.

The problem is to identify these central configurations. For three particles on a line, the directional requirement is already satis-

fied, and the magnitudes of the forces can be balanced by shifting the middle particle relative to the other two. If the end particles have equal masses, for instance, the middle particle must be at the midpoint; if unequal, the middle particle must be closer to the lighter outer one.

But in general, with three particles that aren't aligned (as in Fig. 1 on page 51), directional requirements introduce new complications. Two of them must attract the third towards the centre of mass, therefore they must be placed so that forces perpendicular to that direction cancel. This symmetry requirement must be simultaneously satisfied for the positioning of each pair relative to the remaining particle. Surprisingly, for any choice of masses, the three-body central configuration for Newtonian gravity is simply an equilateral triangle. (This is illustrated by the Trojan asteroids, which follow Jupiter's orbit around the Sun but are offset by 60°.) But as the number of bodies increases, it becomes much harder to work out the necessary balancing act, and we know little about the central configurations of four or more bodies.

Buck has stepped around this problem, not by cutting a Gordian knot, but by tying some knots of his own. His closed-loop solutions must satisfy a similar condition to that in the central configuration problem, somehow eliminating any forces that would pull the particles away from the fixed orbit. Again, this imposes powerful constraints.

For gravity to force a particle around each bend of a curve, the attracting force has to reflect the properties of the bend. Here it is fortunate that Newton's inverse-square law requires the closest particles to have the strongest effect. So, if the particles are packed closely enough together, the net force, dominated by the attraction of nearby particles, is perpendicular to the curve and its magnitude is proportional to the curvature. So if there are many particles of equal size, moving along a curve at constant speed, the acceleration they experience due to gravity is almost exactly right to pull them around each bend in the curve. One gets an approximate n -body solution simply by constructing any closed curve.

The study of problems with an arbitrarily large number of bodies goes back at least to 1856, and Maxwell's study⁴ of the rings of Saturn. He started with a large central body and n small equal-mass particles in a ring. If these particles are spaced equally along a circle and have appropriate initial velocities, they will circle the central mass like a rigid body. To satisfy stability conditions, which allow nearby orbits to behave in a similar fashion, Maxwell introduces an unlimited number of particles.

But there is an important difference between the approaches taken by Maxwell and Buck. Maxwell takes the limit of actual,

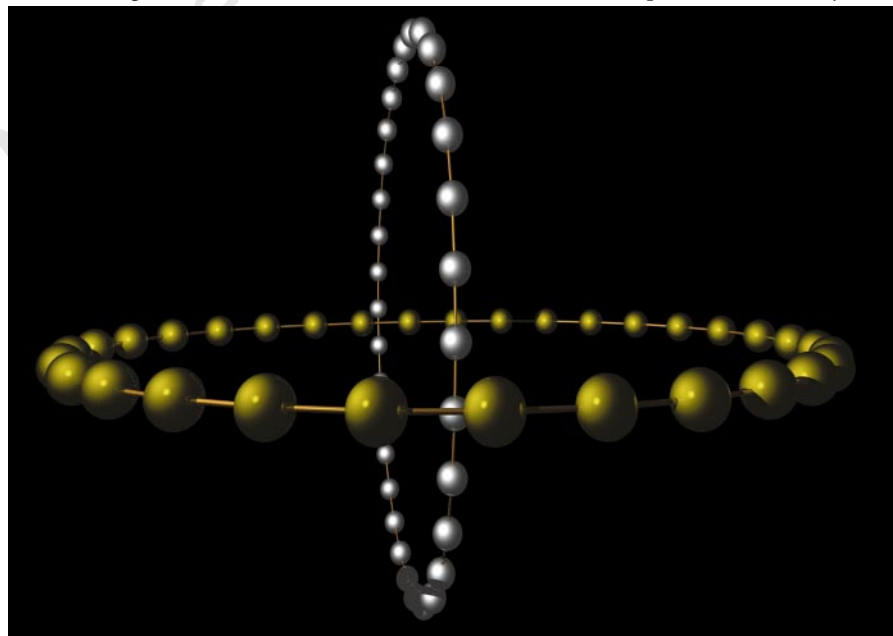


Figure 1 Gravity's chain-link. These strings of bodies can orbit along the closed loops that they define, moving at constant speed. This is just one of a vast new array of approximate solutions to the n -body problem, including any closed curve or set of curves with no intersections. The solutions become more exact as the number of bodies increases, but they may, nevertheless, be unstable.

exact n -body solutions — for each n , the desired solution exists. Buck's conditions, however, need be satisfied only in the limit of $n \rightarrow \infty$; they need not hold for any finite n . Consequently, Buck has to argue that, for large enough n the motion approximates the limiting motion. But small forces remain which would pull particles off the curve. So how long will it take for these contrary forces to destroy the looped orbits? From what we know about the stability of Newtonian mechanics, the orbits might persist for a long time, but they are more likely to deteriorate rapidly. This is a crucial question, and a difficult one to answer.

So we may never see a real example of Buck's looped solutions. Yet, as the history of celestial mechanics has proved, there are ways to convert approximate solutions into useful tools. For instance, the mathematical definition of collisions requires point masses to meet at the same place at the same time. In a sense, each particle eventually treats its colliding partners as clumped in one spot, so we might expect them to align in a manner similar to that required by the central configuration constraints. That indeed happens: for a complete gravitational collapse⁴ and for any collision^{5,6}, the configuration formed by the colliding particles approximates a central configuration. By exploiting this approximation, much of what we know about the mathematics of collisions has been obtained.

Similarly, should particles escape from any configuration with close to the lowest possible escape velocity, they will end up far from the main cluster, travelling slowly. We might expect that the competing attracting forces of other particles will eventually cancel to create a force similar to that in a two-body equation. This approximation to a central configuration has been used to derive the general evolution properties of Newtonian

n -body systems^{6,7}. For example, expanding clusters of galaxies should slowly form central configurations, while the clusters separate from each other at a more rapid rate. In this spirit, we must work out how best to exploit Buck's idea to improve our understanding of the n -body problem. □

Donald G. Saari is in the Department of Mathematics, Northwestern University, Evanston, Illinois 60208-2730, USA.

e-mail: dsaari@nwu.edu

1. Diacu, F. & Holmes, P. *Celestial Encounters* (Princeton Univ. Press, NJ, 1996).
2. Barrow-Green, J. *Poincaré and the Three Body Problem* (Am. Math. Soc., Providence, RI, 1997).
3. Buck, G. *Nature* **395**, 51–53 (1998).
4. Wintner, A. *The Analytic Foundations of Celestial Mechanics* (Princeton Univ. Press, NJ, 1941).
5. Saari, D. G. *J. Differential Equations* **55**, 300–329 (1984).
6. Saari, D. G. *Trans. Am. Math. Soc.* **156**, 219–240 (1971).
7. Marchal, C. & Saari, D. G. *J. Differential Equations* **20**, 150–186 (1976).

Neurobiology

Columns, slabs and pinwheels

Nigel W. Daw

For over 30 years neurobiologists have been intrigued by how the outer layer of the brain — the cerebral cortex — is organized into columns of cells. These develop mainly after birth, under the control of activity driven by the sensory input. There are many models for this development, and, on page 73 of this issue, Wolf and Geisel¹ argue that they are all subject to symmetry principles. These principles lead to some general conclusions about the ways in which the columns develop, independent of when and how development occurs.

The columnar organization of the cerebral cortex was first demonstrated by Mountcastle² in the somatosensory cortex, then elaborated by Hubel and Wiesel³ in the visual cortex. They discovered that cells in a column, extending from the surface of the cortex to the white matter directly underneath, have similar properties. Working with cat visual cortex, Hubel and Wiesel showed that two properties are organized in a columnar fashion — ocular dominance (response to the right or left eyes, or both) and orientation selectivity (selection for the orientation

of a bar; horizontal, vertical or some angle in-between). Because almost every cell has these two properties, there are two overlapping columnar maps. When the macaque visual cortex was investigated, a third property was included — colour. This inserted into the ocular-dominance and orientation-selectivity maps, with a pattern of blobs recognizable in cytochrome-oxidase stains centred on the ocular dominance columns.

Hubel and Wiesel⁴ went on to show that recordings made from long, oblique penetrations of an electrode into the visual cortex yield a steady change in the orientation selectivity of the cells. Sometimes the progression is clockwise, sometimes anticlockwise, and occasionally it is interrupted by a sudden shift in orientation. They suggested that orientation is organized in a series of parallel slabs (the ice-cube model). Later, Braitenberg and Braitenberg⁵ argued that the evidence was better explained by a set of orientation domains radiating out from centres (the pinwheel model). The development of a technique for imaging a large area of the cortex and working out the orientation sensitiv-

Meteors

Things that go bump in the night sky

Every 33 years, or thereabouts, the Leonid meteor shower is spectacular — so many meteors fall that it is called a storm. The last time was in 1966, when up to 40 meteors a second were seen; and back in 1833, watchers even heard sounds. That could be telling us about processes occurring on comet 55P/Tempel–Tuttle (shown here), the creator of the Leonids.

As comets are heated by sunlight, their ices sublime. The breeze of this sublimation lifts debris off the comet nucleus, and some of these pieces eventually get swept up by the Earth. We see them as meteors — usually just a streak of light, created by a body a few millimetres in diameter. Larger bodies produce brighter fireballs.

Sounds are much rarer, but in 1833



many observers heard hissing, crackling and popping noises. These were probably 'electrophonic sounds', created by very-low-frequency radio waves, in turn generated by turbulent plasma in the wake of a disintegrating meteoroid. According to

Martin Beech (*Astron. J.* **116**, 499–502; 1998), the falling bodies must have been well over a metre across to produce such sounds.

That is much too big to have been lifted off the nucleus of Tempel–Tuttle by sublimation pressure. So how did these great lumps of dirt escape the comet? Perhaps there are cavities in the nucleus that are only occasionally lit by the Sun, producing violent outbursts that can shoot out large bodies.

We may be in for another big show from the Leonids this year, around 17 November. Then, it may even be possible for astronomers to see these large meteoroids directly, or to see them hitting the Moon. And if you're watching, listen out.

Stephen Battersby

TIM PUCKETT

ity for a series of different orientations in the same animal allowed Bonhoeffer and Grinvald⁶ (using intrinsic signals of activity from upper layers of cortex⁷) to show that the pinwheel model is correct.

There are many models of how the cortex develops to achieve this pattern⁸. Cats and macaques show a pattern of orientation selectivity as soon as the eyes open shortly after (or at) birth. But ferrets, which are born at an earlier stage of development, show little orientation selectivity when their eyes open. Instead, orientation selectivity develops to adult levels during the fifth and sixth weeks after birth, under the influence of nerve-cell activity⁹.

Wolf and Geisel¹ now suggest that all models for the development of orientation selectivity are subject to symmetry principles. The authors first assume that columns with a similar orientation preference tend to be a particular distance apart. Their second assumption is that patterns resulting from (i) shifting the pattern parallel to the cortical surface; (ii) rotating the pattern parallel to the cortical surface; and (iii) shifting all orientation preferences by the same angle, will arise with the same probability. Their final assumption is that the pattern arises from many random factors, and that Gaussian statistics apply. Their mathematics then shows that the expected density of pinwheels is given by

$$\rho = \frac{\pi(1+\alpha)}{\Lambda^2}$$

where Λ is the average spacing of columns, and α is a number that describes the structure of spatial correlations. One therefore expects to find at least π pinwheels in an area of size Λ^2 .

In fact, the density of pinwheels varies from 3.75 in an area of size Λ^2 in macaque, to 2.1–2.6 in tree shrew, with intermediate values for squirrel monkey, ferret and cat. Thus, either there is some process that prevents development from being carried to completion in the cat and tree shrew, where the value is substantially below π , or the density of pinwheels develops to a value greater than π , then decreases through annihilation of pinwheels. Wolf and Geisel propose that such an annihilation could occur through melding of a pair of pinwheels — one with clockwise chirality and another with anticlockwise chirality.

The annihilation of pinwheels is an intriguing possibility. But it is not supported by the few observations available, which show a series of orientation maps that does not change much with age⁸. Areas with longer stretches of parallel slabs — which are predicted to arise when two pinwheels merge — do not seem to appear much more frequently in the older patterns. So, Wolf and Geisel's predictions still need to be rigorously tested. Moreover, the mathematics

assumes that there are no intrinsic places in the cortex that would attract the formation of pinwheels. This has been suggested as a factor in positioning of the blobs for colour in the macaque visual cortex¹⁰.

Wolf and Geisel have used mathematical approaches that are common in physics but rare in biology. In turbulent flow, for example, aspects of a system's behaviour can be predicted from basic symmetries without knowing the detailed nature of the interactions between components of the system. Because the nervous system is made up of billions of cells and trillions of synapses, with detailed interactions that are unknowable in practice, mathematical approaches may have a future in predicting the overall behaviour of the system. □

RNA processing

A tale of two tails

David Bentley

The fate of RNAs in the nucleus is determined by the polymerase that made them. For example, the same primary transcript that will give rise to a messenger RNA if made by RNA polymerase II (pol II), will never mature if it is made by pol I, pol III or a bacteriophage RNA polymerase^{1,2}. How can we explain this link between the machines that make and process mRNAs? On page 93 of this issue, Hirose and Manley³ bring us one step closer to an answer. They report that the carboxy-terminal domain (CTD) — a protein domain unique to pol II — acts as part of the protein complex that fashions the 3' end of the mature mRNA by cleavage and polyadenylation.

The formation of 3' ends is a two-step process in which the RNA transcript is cleaved downstream of the sequence AAUAAA, then a poly(A) tail is added at the cut end. RNAs made *in vivo* by pol II lacking the CTD — a conserved repeat of the

Nigel W. Daw is in the Department of Ophthalmology and Visual Science, Yale University School of Medicine, 330 Cedar Street, New Haven, Connecticut 06520-8061, USA.

e-mail: Daw@biomed.med.yale.edu

1. Wolf, F. & Geisel, T. *Nature* **395**, 73–78 (1998).
2. Mountcastle, V. B. *J. Neurophysiol.* **20**, 408–434 (1957).
3. Hubel, D. H. & Wiesel, T. N. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
4. Hubel, D. H. & Wiesel, T. N. *J. Comp. Neurol.* **158**, 267–294 (1974).
5. Braitenberg, V. & Braitenberg, C. *Biol. Cybern.* **33**, 179–188 (1979).
6. Bonhoeffer, T. & Grinvald, A. *Nature* **353**, 429–431 (1991).
7. Grinvald, A., Lieke, E. E., Frostig, R. D., Gilbert, C. D. & Wiesel, T. N. *Nature* **324**, 361–364 (1986).
8. Erwin, E., Obermayer, K. & Schulten, K. *Neural Comp.* **7**, 425–468 (1995).
9. Chapman, B., Stryker, M. P. & Bonhoeffer, T. *J. Neurosci.* **16**, 6443–6453 (1996).
10. Kuljis, R. O. & Rakic, P. *Proc. Natl Acad. Sci. USA* **87**, 5303–5306 (1990).

sequence YSPTSPS, found at the carboxy terminus of the pol II large subunit — are not efficiently capped, spliced or cleaved at the poly(A) site^{4,5}. This suggests that the CTD helps to target mRNA processing factors to pol II transcripts^{6,7}. Hirose and Manley have now discovered that the CTD facilitates the 3' RNA cleavage reaction, even in the absence of transcription. They suggest that, in effect, the CTD is a cofactor for 3' processing.

Production of a mature mRNA 3' end can be reconstituted *in vitro* with cleavage/polyadenylation specificity factor (CPSF), cleavage-stimulation factor (CstF), cleavage factors CFI and CFII, and poly(A) polymerase (PAP)⁸. The association of CPSF and CstF with the CTD *in vitro*⁴, and the colocalization of CstF and phosphorylated pol II *in vivo* (Fig. 1), indicate that a stable 'mRNA factory' complex of pol II with processing factors carries out synthesis and

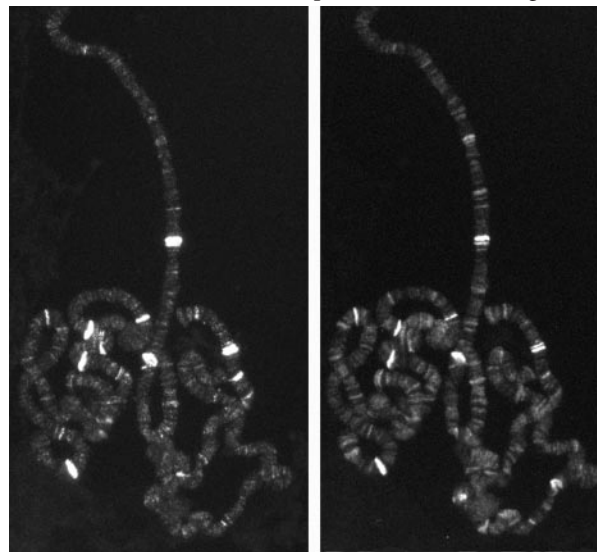


Figure 1 Colocalization of the cleavage-stimulation factor CstF and phosphorylated RNA polymerase II (pol II). Immunofluorescence of *Drosophila* polytene chromosomes with anti-pol II monoclonal antibody (left) and rabbit antibody against Suppressor of forked, the *Drosophila* homologue of CstF (right). (Courtesy of M. Sikes and A. Beyer, University of Virginia.)

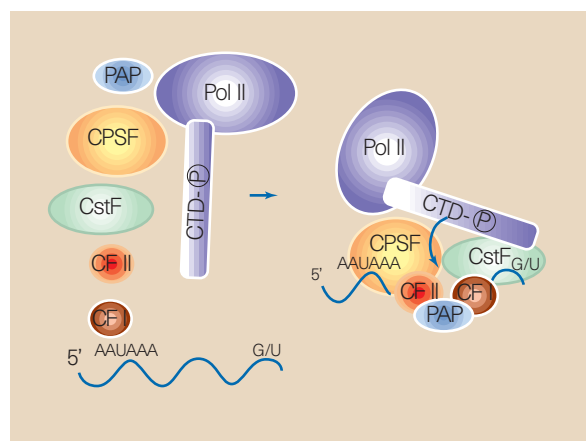


Figure 2 Model for activation of 3' processing by the pol II carboxy-terminal domain (CTD). Hirose and Manley³ have shown that the CTD is part of the protein complex that carries out cleavage and polyadenylation to produce a mature messenger RNA transcript. CPSF, cleavage/polyadenylation specificity factor; CstF, cleavage-stimulation factor; CFI, cleavage factor I; CFII, cleavage factor II; PAP, poly(A) polymerase.

maturation of the primary transcript⁹.

Hirose and Manley provide an insight into how the mRNA factory operates. They followed up the curious observation that, in a purified system, cleavage at the poly(A) site is stimulated by creatine phosphate and by phosphoamino acids. They reasoned that these compounds may mimic a phosphoprotein that allosterically activates the cleavage reaction. An attractive candidate phosphoprotein is pol II — the CTD undergoes a cycle of hyperphosphorylation and dephosphorylation (mainly at residues two and five of the YSPTSPS repeat¹⁰) as pol II cycles through transcriptional initiation, elongation and termination. Sure enough, the authors found that phosphorylated pol II and recombinant CTD stimulate cleavage *in vitro*. Unexpectedly, however, the unphosphorylated CTD was also effective.

These experiments show that the CTD stimulates 3' processing in the absence of transcription, although it is not clear whether the CTD is acting in the same way as creatine phosphate. Hirose and Manley also found that the cleavage reaction was inhibited when they immunodepleted pol II from a crude nuclear extract, after dissociating poly(A) factors from the CTD with high salt. Moreover, they could restore cleavage activity by adding back pure pol II. In other words, the CTD stimulates cleavage in a crude extract, as well as in a purified system.

But how does the CTD stimulate cleavage in the absence of transcription? Does it promote the formation of a stable complex (within which cleavage can then occur), or does it contact polyadenylation factors only fleetingly, triggering cleavage? Assembly of a complex — which might comprise CstF, CPSF, CFI, CFII and PAP, together with the RNA (Fig. 2) — may be rate limiting both *in vivo* and *in vitro*. The CTD could accelerate assembly of the complex by acting as a scaffold that positions the proteins optimally, all in the same place at the same time. A scaffold might even impose an order of assembly of the polyadenylation machine.

Hirose and Manley suggest that the CTD is not just a passive scaffold but an active participant in the cleavage reaction. According

to this model, which is more consistent with the effect of creatine phosphate, the CTD behaves as a cofactor or allosteric activator. One possible target for allosteric activation is CstF, which binds the CTD *in vitro*⁴. The importance of this interaction for scaffolding or allosteric activation could be tested using CTD mutants that do not bind CstF.

It is now clear that pol II stimulates 3' processing, both *in vivo* and *in vitro*, through the CTD. Conversely, poly(A) factors may affect pol II by influencing the decision to terminate or elongate the RNA chain. Termination depends on transcription of the poly(A) signal¹¹, poly(A) factors¹² and the CTD (ref. 4). Signalling from the polyadenylation machine to the polymerase may ensure that the chain is terminated only after pol II has reached the end of a gene. So, a two-way line of communication between pol II and the 3' processing factors may be built into the mRNA factory.

Working out how the machines in the mRNA factory interact with one another is a big challenge. Hirose and Manley³ have provided the first *in vitro* system to study one aspect of the complex communication system that integrates different nuclear processes through protein–protein interactions. We can look forward to the development of more *in vitro* systems, which will reveal the details of how the CTD affects splicing and capping as well as polyadenylation. □

David Bentley is in the Department of Biochemistry and Molecular Genetics, UCHSC, 4200 East 9th Avenue, Denver, Colorado 80262, USA.
e-mail: David.Bentley@UCHSC.edu

1. Sisodia, S. S., Sollner, W. B. & Cleveland, D. W. *Mol. Cell. Biol.* **7**, 3602–3612 (1987).
2. Smale, S. T. & Tjian, R. *Mol. Cell. Biol.* **5**, 352–362 (1985).
3. Hirose, Y. & Manley, J. L. *Nature* **395**, 93–96 (1998).
4. McCracken, S. et al. *Nature* **385**, 357–361 (1997).
5. McCracken, S. et al. *Genes Dev.* **11**, 3306–3318 (1997).
6. Greenleaf, A. L. *Trends Biochem. Sci.* **18**, 117–119 (1993).
7. Yuryev, A. et al. *Proc. Natl Acad. Sci. USA* **93**, 6975–6980 (1996).
8. Keller, W. & Minvielle-Sebastia, L. *Curr. Opin. Cell Biol.* **9**, 329–336 (1997).
9. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. *EMBO J.* **12**, 1059–1065 (1993).
10. Dahmus, M. E. *J. Biol. Chem.* **271**, 19009–19012 (1996).
11. Logan, J., Falck-Pederson, E., Darnell, J. & Shenk, T. *Proc. Natl Acad. Sci. USA* **84**, 8306–8310 (1987).
12. Birse, C., Minvielle-Sebastia, L., Lee, B., Keller, W. & Proudfoot, N. *Science* **280**, 298–301 (1998).

Daedalus

Zero-tolerance policing

Millions of potentially lethal cell mutations occur in each of us every day. We notice nothing; the immune system arrests and executes the offenders instantly. But with advancing age, its vigilance slackens. Mutations start to go unchecked; and we succumb to the melancholy neoplastic or degenerative diseases of old age.

So Daedalus is extending his scheme of last week for priming the immune system against all normal biological proteins. He now wants to prime it against abnormal proteins as well — in particular, the abnormal ones expressed on the surfaces of malfunctioning human cells.

LEADCO volunteers are now providing small biopsy tissue samples of blood, lung, gut, and so on, to be cultured in the usual way. The cultures will then be attacked with mutagens, ultraviolet light, X-rays, and such, to induce as many mutations in them as possible. The battered and misfiring cells will express all sorts of crazy and altered proteins on their surfaces. These will be extracted, and injected back into the volunteer from whom the cells came. This 'auto-immunization' will prime his immune system against such cell mutations. Should a cell in his gut or liver start to go crazy in one of those pre-identified ways, the immune system will spot the abnormality at once. It will stamp on the rogue cell instantly.

Sadly, auto-immunization can never be a perfect defence. There are too many different body tissues, each with too many ways of going wrong, for any vaccine to contain all the proteins expressed by all these errors. But not all errors are equally likely, or equally deadly if they do occur. The human genome can suffer about 10 billion possible single-site mutations; yet there are only some 30,000 recognized illnesses of all kinds. The deadly diseases are those in which a mutated cell starts to multiply, or to trigger similar changes in its neighbours. So Daedalus will tune his treatment to his customers. If all your ancestors died of liver cancers, you'll include some liver cells in your tissue sample; if Dr Alzheimer stalks your family tree, you'll make sure that brain cells are well represented in it. And the mutagenic regime itself will be chosen to provoke the most appropriate type of cell damage. The resulting immunization will be an excellent defence against the disease of your fears. Sadly, it won't make you immortal. But you'll last much longer, and finally die of something you weren't expecting.

David Jones

Heezen's highlands

People who chart the ocean floor draw up landscapes no one has seen, using machines that send out sound waves and invisible rays. Turning sound into shape, their achievement is a map we can see and understand.

Martin Kemp

Systematic mapping of the ocean floor has progressed, in its relatively short existence, through stages that mirror in important respects the history of terrestrial maps from the time of the Renaissance. Yet the prime early means for representing undersea topographies involved emissions not discernible by our eyes.

The technical development of echosounding was driven by submarine warfare. Using this sonar technique, pioneering oceanographers, including Maurice Ewing, Bruce Heezen and Marie Tharp, turned their attention to the remarkable highlands and lowlands that were being progressively disclosed in the depths of the oceans.

The earliest charts or 'physiographic diagrams', made possible by the increasingly high resolution of echo-sounders in the 1950s, synthesized exactly the same components of observed measurement, logical extrapolation, approximation, theory and imaginative guesswork that had characterized the mapping of *terra incognita* in the wake of Renaissance voyages of discovery.

The basic data consisted of the linear tracks of the surveying lines, often quite widely spaced, and the sonar measurements of the profiles of the ocean floor along these lines.

Tharp has described how she and Heezen conceived the first coherent relief map of the Atlantic in 1952, and how Heezen dashed off his first inspirational sketch: "he seized a piece of paper and within an hour or so drew in the topography". On the basis that "the proper way is to present it as it would be seen if all the water were drained away", they came to operate with three principles.

First, "hypotheses of ocean floor structure must be used to supplement the often meagre data, and only the use of correct hypotheses will result in maps closely approximating nature". **Second, "what data exist in several disciplines must be put at one scale".** **Third, "sketching proceeds from the shoreline seawards and then from the mid-ocean crest landwards, as the policy was to go from the known to the unknown".**

The resulting maps were greatly refined in data and technique over the subsequent years — involving, among others, Heinrich Berann, the Austrian painter of Alpine panoramas. They notably resemble historic maps in fictive relief, as exemplified by the sixteenth-century painted topographies



Heezen drew in the topography within an hour or so for the first physiographic sketch of the western North Atlantic; from *The Ocean Floor*, ed. R. Scrutton and M. Talwani (Wiley, 1982).

designed by Ignatio Danti (Palazzo Vecchio, Florence, and the Vatican) and Paul and Thomas Sandby's eighteenth-century surveys of the unruly Highlands of Scotland.

Heezen and Tharp's results, published in the form of both technical diagrams and popular panoramas, satisfy ancient and modern pictorial impulses in equal measure. Reflected sound at the frequency of bird-song is translated into pictorial images in a

way that has become characteristic of recent scanning devices that use such non-visible emissions as X-rays and positrons. It seems that our pictorial proclivities refuse to go away, however non-sensory the techniques of machine perception become. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

e-mail: martin.kemp@trinity.oxford.ac.uk



From sketch to detail: profiles help in the building of the physiographic panorama of the Atlantic; from *The Face of the Deep*, by Bruce C. Heezen and Charles D. Hollister (Oxford University Press, 1971).

Promiscuity in transgenic plants

The ecological risks of genetically modified crops are of greatest concern when there are no inherent barriers to the spread of transgenes through sexual reproduction. This is most likely when transgenes can spread to weedy species through hybridization, or when the crop species itself exists in weedy forms¹. If the potential recipient of a transgene is a highly selfing species, such as *Arabidopsis thaliana*, this risk is often considered negligible². Here, however, we report results of a field experiment in which transgenic *A. thaliana* showed a dramatically increased ability to donate pollen to nearby wild-type mothers compared with *A. thaliana* mutants expressing the same mutant allele as the transgenic plants.

In the summer of 1996, we planted 144 *A. thaliana* rosettes at random locations in a grid at our field site in central Illinois, at densities reflecting those of nearby *A. thaliana* populations. One-quarter of our plants were homozygous for the dominant and mutant allele of acetolactate synthase, *Csr1-1*; this allele confers resistance to the herbicide chlorsulphuron. Mutants were originally isolated through mutation of the wild-type *A. thaliana* ecotype, Columbia, by ethyl methanesulphonate, and had been backcrossed for six generations³. One-quarter of the plants in the field were wild-type Columbia, and the remaining half were Columbia-strain plants transformed to express *Csr1-1* in a pBin vector^{4,5}. The latter possessed insertions at a single site, were homozygous for this insertion, and were equally divided between two independently transformed lines.

Plants were grown in the absence of herbicides, and were visited frequently by syrphid flies which consume pollen and nectar. At the season's end, we identified

progeny produced through outcrossing by germinating seeds produced by each wild-type Columbia plant on plates containing 100 nM chlorsulphuron. Chlorsulphuron-resistant seedlings were transplanted to the greenhouse, where outcrossing was prevented with pollination bags, and their selfed seeds were collected. We then identified progeny fathered by transgenic *A. thaliana* by germinating selfed seeds on plates containing 50 mg l⁻¹ kanamycin (only transgenic fathers were resistant to both chlorsulphuron and kanamycin).

A survey of approximately 100,000 seeds showed that the per-plant outcrossing rate was 0.30% for mutant fathers and 5.98% for transgenic fathers. Transgenic *A. thaliana* were roughly 20 times more likely to outcross than ordinary mutants. We screened a subsample of 281 transgenic progeny with primers specific to each insertion site in order to identify which of the transgenic lines had fathered them. We calculated the outcrossing rates for these two lines as 1.2% and 10.8%. Both transgenic lines showed increased outcrossing relative to the mutant (χ^2 tests, $P < 0.003$ in each case), but the two transgenic lines differed in their propensity to outcross ($P < 0.001$).

To be certain that seed contamination did not contribute to estimates of outcrossing, we re-plated the seeds from all selfed, resistant plants on chlorsulphuron plates. Contaminants would have been homozygous for resistance whereas progeny resulting from outcrossing would have been heterozygous. Over 99% of our resistant progeny were heterozygous, and only heterozygous individuals were included in our calculations. We also placed known resistant and susceptible seeds on all selection

plates to confirm our ability to determine the phenotypes of progeny.

Our results show that wild-type *A. thaliana* are more likely to be fertilized by the pollen of transgenic rather than mutant *A. thaliana* when each expresses the mutant allele *Csr1-1*. Although *A. thaliana* is unlikely to become a pernicious weed, these results show that genetic engineering can substantially increase the probability of transgene escape, even in a species considered to be almost completely selfing.

Our results do not prove that enhanced outcrossing is due to the transgene itself, but rather that a difference in outcrossing between transgenic and mutant plants exists. Because we do not know the underlying genetic mechanism, the generality of our result is unclear at present. Even if enhanced outcrossing is restricted to *Csr1-1*, however, our results are of broad relevance because this transgene has been introduced into dozens of agricultural crops, and is advocated as a selectable marker for plant transformation vectors⁶.

Joy Bergelson, Colin B. Purrington*, Gale Wichmann

Department of Ecology and Evolution,
University of Chicago, 1101 East 57th Street,
Chicago, Illinois 60637, USA

e-mail: jbergels@midway.uchicago.edu

*Present address: Department of Biology,
Swarthmore College, 500 College Avenue,
Swarthmore, Pennsylvania 19081, USA

1. Mikkelsen, T. R., Andersen, B. & Jorgensen, R. B. *Nature* **380**, 31 (1996).
2. Tiedje, J. M. et al. *Ecology* **70**, 298–315 (1989).
3. Haughn, G. W. & Somerville, C. *Mol. Gen. Genet.* **204**, 430–434 (1986).
4. Bergelson, J., Purrington, C. B., Palm, C. J. & López-Gutiérrez, J.-C. *Proc. R. Soc. Lond. B.* **263**, 1659–1663 (1996).
5. Frisch, D. A. et al. *Plant Mol. Biol.* **27**, 405–409 (1995).
6. Li, Z., Hayashimoto, A. & Mural, N. *Plant Physiol.* **100**, 662–668 (1992).

Resistance to the herbicide glyphosate

The News and Views article by Gray and Raybould¹ paints a rosy picture of the future of agricultural crops engineered with resistance to the herbicide glyphosate. It is true that the presence of the glyphosate resistance transgene (the EPSPS gene) in the chloroplast genome of crop plants would eliminate pollen transfer of this gene to weedy relatives. But there have already been large-scale releases of transgenic, glyphosate-resistant ('Roundup Ready') crops, including glyphosate-resistant *Brassica napus* (Canola, varieties Quest and LG3295), which is disturbing as there are

many *Brassica* weeds that could potentially hybridize with the engineered *B. napus* (for example, see ref. 2).

Gray and Raybould¹ also say that "[glyphosate] tolerance in weeds does not seem to have evolved, despite extensive use of the herbicide for over 20 years". Although accurate about the extent of glyphosate use, this statement is inaccurate about the occurrence of spontaneous glyphosate resistance. There have been at least two reported cases of glyphosate resistance in *Lolium rigidum* (rigid ryegrass) in Australia. An article by scientists at Monsanto³ cited by Gray and Raybould mentions one of these, stating that "studies are required to determine the basis of resistance". Monsanto, who produce a commercially available glyphosate ('Roundup'), are collaborating

with a group at Charles Sturt University in New South Wales, Australia, to understand the nature of the resistance in this case. More information is available from a transcript of an ABC Radio National discussion (<http://www.abc.net.au/rn/talks/bbing/bb970914.htm>) and a catalogue of herbicide resistance (<http://www.weedscience.com>).

It seems foolish to assume that resistance to a herbicide with a single biochemical target *in vivo* would not develop, given time. Yet public concern about the ability of plants to develop resistance to glyphosate has been assuaged. Perhaps this has caused many other 'isolated' incidences of spontaneous glyphosate resistance to go unnoticed. We should all be concerned about the presumed increased use of glyphosate arising from the cultivation of transgenic

glyphosate-resistant crops. It might be time to demote glyphosate from its status as a completely safe, fix-all herbicide.

Stanley Robert

Department of Zoology,
University of Washington,
Box 351800, Seattle, Washington 98195, USA
e-mail: srobert@u.washington.edu

Ute Baumann

Department of Plant Science,
University of Adelaide,
Glen Osmond, S.A. 5064, Australia

- Gray, A. J. & Raybould, A. F. *Nature* **392**, 653–654 (1998).
- Landbo, L. & Jorgensen, R. B. *Euphytica* **97**, 209–216 (1997).
- Bradshaw, L. D., Padgett, S. R., Kimball, S. L. & Wells, B. H. *Weed Technol.* **11**, 189–198 (1997).

Turning off follicular dendritic cells

Follicular dendritic cells (FDCs) are specialized to display intact antigens for recognition by B cells. Here we show that inhibition of lymphotoxin α/β (LT α/β) leads to the disappearance of multiple markers on FDCs within one day, regardless of whether any immune reactions are taking place. Inhibition of the tumour necrosis factor (TNF) pathway was also effective, but only in the absence of a strong antigenic response. These treatments not only prevented trapping of newly formed immune complexes, but also eliminated previously trapped immune complexes on FDCs, rendering the FDC non-functional. Therefore lymphotoxin and/or TNF is required for the continual maintenance of FDC function.

Gene deletion experiments in mice have shown that signalling by both TNF and lymphotoxin is required for FDC development^{1–3}. TNF and lymphotoxin make up two separate immunoregulatory pathways⁴. TNF and potentially the soluble LT α form signal through the classic TNF receptor (TNF-R)-mediated pathway. Membrane-bound LT α/β heteromers signal through the LT β receptor (LT β R)⁵, forming a second pathway associated with the development and maintenance of secondary lymphoid organs^{1,6}.

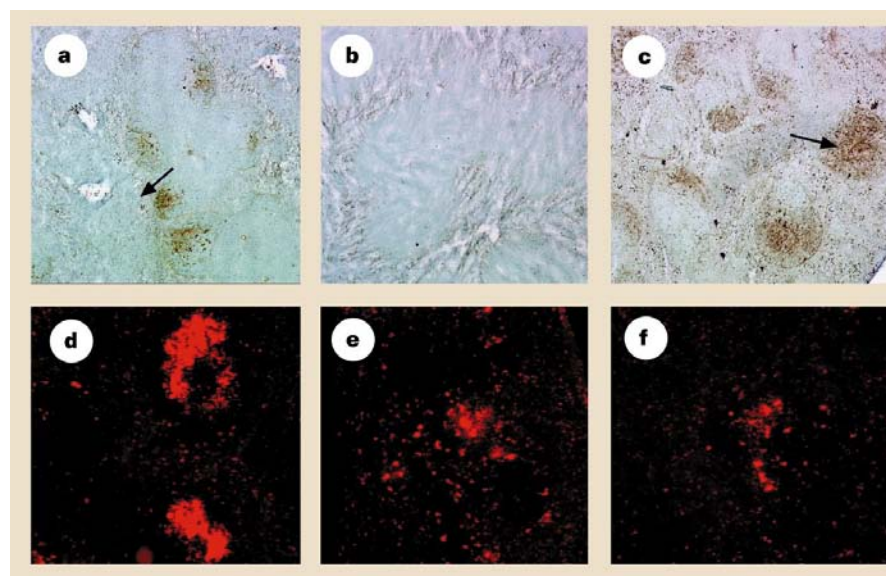


Figure 1 Only the LT α/β pathway is essential for the maintenance of FDC networks following immunization with sheep red blood cells, yet both the TNF and lymphotoxin pathways are required for the maintenance of immune complexes. Mice were treated as previously described⁶ with control human limmunoglobulin (**a**, **d**), LT β R-Ig (**b**, **e**) or TNFR55-Ig (**c**, **f**). In **a**–**c**, animals were immunized with sheep red blood cells, and frozen sections of the spleen were analysed by immunohistochemistry for the presence of the FDC marker FDC-M1. In **d**–**f**, mice were immunized twice with phycoerythrin, and frozen sections from the spleen collected 8 hours after the last immunization were observed under the fluorescence microscope.

To determine the roles of TNF, LT α/β and CD40L in maintaining mature networks of FDCs, we examined the spleens of normal adult mice treated with LT β R-Ig or anti-LT β antibody (BB-F6) to block the LT α/β pathway, TNF-R55-Ig to inhibit the TNF/LT α pathway, or anti-CD40L (MR1) to inhibit the CD40L pathway (Table 1). Inhibition of either the TNF or LT pathway in non-immunized mice led within 24 h to the disappearance of the FDC-specific markers FDC-M1, FDC-M2, MAdCAM-1 and CR1. The same observation was made in lymph nodes. The elimination of MAdCAM-1 expression on reticular cells in the marginal zone also tracked with the FDC, whereas the disappearance of markers for marginal-zone macrophages took weeks⁶, raising the possibility that these MAdCAM-1-positive cells are FDC precursors. The effect of LT α/β pathway inhibitors was seen with and without immunization with keyhole limpet haemocyanin or sheep red

blood cells, and treatment with TNF-R55-Ig had no effect when the animals were immunized (Fig. 1a–c and Table 1).

We investigated the function of FDCs by testing their capacity to trap immune complexes in the spleen⁷. Immune complexes were formed following a secondary immunization with phycoerythrin, and the fluorescent property of the protein allowed immune complexes to be detected in B-cell follicles. These immune complexes are known to colocalize with FDCs⁷. Pretreatment of mice with inhibitors to the LT α/β pathway before immunization with phycoerythrin prevented trapping of immune complexes on FDCs (data not shown). Pretreatment with TNF-R55-Ig did not prevent immune complex trapping, although the pattern of immune complex deposition was altered. Once immune complexes are trapped on FDCs, inhibition of either the LT α/β or TNF pathway now affected the maintenance of these immune complexes (Fig. 1d–f and Table 1). Moreover, because anti-LT β had the same effect as LT β R-Ig in all cases, involvement of the newly defined LIGHT ligand can be excluded in these functions⁸. Treatment with anti-CD40L antibody had no effect in any of the experiments.

Pretreatment of adult mice with inhibitors of the LT α/β pathway impairs the antibody response to sheep red blood cells⁶. We used phycoerythrin-specific enzyme-linked immunosorbent assays to show that treatment with LT β R-Ig or TNF-R55-Ig did not inhibit the primary or the secondary antibody response to phycoerythrin. Thus,

Table 1 How TNF and lymphotoxin pathways affect FDC status in the spleen

Treatment	Control Ig	LT β R-Ig or anti-LT β	TNFR55-Ig	Anti-CD40L
Detection of FDC in spleen after immunization with				
No antigen	Yes	No	No	Yes
Sheep red blood cells	Yes	No	Yes	Yes
KLH	Yes	No	Yes	Yes
Phycoerythrin*	Yes	No	No	Yes
Deposition of immune complex on FDCs				
	Yes	No	Yes	Yes
Maintenance of immune complex on FDCs†				
24 h	Normal	Normal	Normal	Normal
48 h	Normal	Reduced	Reduced	Normal
72 h	Normal	Not detected	Not detected	Normal

*Weak antigen.

†Time after treatment of mice previously immunized with phycoerythrin.

the lack of immune complex trapping was not due to the lack of preformed immune complexes. In further support of this finding, passive immune complex trapping following intravenous injections of peroxidase–antiperoxidase immune complexes was not observed in animals treated with LT β R-Ig. We therefore conclude that the trapping defect resulted directly from changes in FDC status or the transfer of immune complexes to the FDC.

Our data show that the TNF and LT α / β axes differ in their effect on the function of a fully developed system. LT α / β signaling seems to be involved in maintaining FDC networks and possibly immune complex transport from the marginal zone. But the role of TNF seems to be more restricted to the maintenance of pre-existing FDC networks. This difference parallels the requirement in adult mice for primarily the LT α / β axis, rather than the TNF pathway, for maintaining the architecture of the splenic marginal zone⁶. A surprising observation is that the maintenance of pre-existing FDCs in a differentiated state requires continual interaction with B lymphocytes expressing LT α / β ⁹. The FDCs can collapse from their mature state within one day when deprived of LT α / β signalling, which is in contrast to the static picture of an FDC network retaining antigen for months to years. Because surface LT α / β is believed to be expressed only on activated lymphocytes, it is not known which cells provide these continuous maintenance signals. Given this dependence on continual LT α / β signalling for maintenance, one could speculate that FDC networks may be dramatically affected by immune suppression, either pharmacological or as a consequence of viral infection.

The FDC compartment has not been amenable to manipulation in normal adult animals, but our technique makes it possible to alter the repertoire of sequestered antigens, allowing investigation of the effect on immunological memory in health and disease. The role of FDCs in immunoglobulin-mediated autoimmune diseases and follicular lymphoma can also be examined. FDC-sequestered HIV particles make up a major viral reservoir in HIV patients¹⁰, and the release of this viral compartment, using inhibitors of the LT α / β pathway, may improve subsequent antiviral treatments.

Fabienne Mackay, Jeffrey L. Browning

Department of Immunology, Inflammation and Cell Biology, Biogen, 12 Cambridge Center, Cambridge, Massachusetts 02142, USA.
e-mail: fabienne_mackay@biogen.com

- Curr. Top. Microbiol. Immunol.* **198**, 175–218 (1995).
6. Mackay, F. *et al.* *Eur. J. Immunol.* **27**, 3033–3042 (1997).
7. Vora, K. A., Ravetch, J. V. & Manser, T. *J. Immunol.* **159**, 2116–2124 (1997).
8. Mauri, D.N. *et al.* *Immunity* **8**, 21–30 (1998).
9. Gonzalez, M. *et al.* *J. Exp. Med.* **187**, 997–1007 (1998).
10. Burton, G.F. *et al.* *Immunol. Rev.* **156**, 185–197 (1997).

Nerve agents degraded by enzymatic foams

The decontamination of large areas that have been exposed to chemical weapons is a critical component of defence and anti-terrorist strategies around the world^{1–3}. Until now, nerve agents could be decontaminated only with bleach treatment and/or *ex situ* incineration⁴, which had severe environmental consequences. Small-scale methods of decontamination and demilitarization are available, but the problem of decontamination over a wide area has not been solved^{2,3}. Here we incorporate organophosphorus hydrolase (OPH), an enzyme that hydrolyses nerve agents, into aqueous fire-fighting foams to catalyse surface decontamination. The performance of enzyme-containing foams is reproducible and predictable in detoxifying organophosphorus nerve toxins deposited onto surfaces. Such foams are an environmentally friendly alternative to current decontamination solutions, which are nonspecific in their action and contain significant amounts of hazardous organic solvents.

Although agent specificity limitations remain, enzyme-catalysed degradation of nerve agents has been demonstrated^{5–7}, with the most widely studied enzyme being the broad-spectrum OPH^{8–12}. Contaminated

surfaces can be treated with aqueous OPH solutions, but the free enzyme in aqueous solution is not sufficiently stable or easy to deliver for wide-area decontamination. Surface decontamination methods may be complicated by the agent becoming more volatile, and by the difficulty of extracting agents from surfaces. When OPH is integrated into fire-fighting foam, however, volatilization is minimized and agents can be extracted from surfaces into the surfactant-containing foam.⁶

Enzymes usually suffer considerable loss of activity when exposed to shear-induced foaming, ionic surfactants and air–water interfaces, but OPH foams can effectively treat glass contaminated with chemical agents. Foams were produced using “First Defense” aqueous fire-fighting foam surfactant solution, a gift from the City of Pittsburgh Fire Department. Fire-fighting foams release water when applied to a surface so it is important to determine how the enzyme will partition between the foam and the discrete aqueous phase. We collected fallen aqueous film from an OPH foam in several fractions over time. We studied enzyme activity in the foams using paraoxon as a model agent. All samples demonstrated identical enzyme activity regardless of collection time. Furthermore, approximately 30% of the enzyme activity originally loaded to the formulation was evenly distributed within the foam and not preferentially lost to the growing liquid phase. Assessment of the kinetic parameters for the enzyme demonstrates that activity loss results directly from a reduction in $V_{\max}$ for the foamed enzyme. Because 30% activity is more than sufficient for use as an effective catalyst, we have examined the effect of enzyme concentration and time of exposure

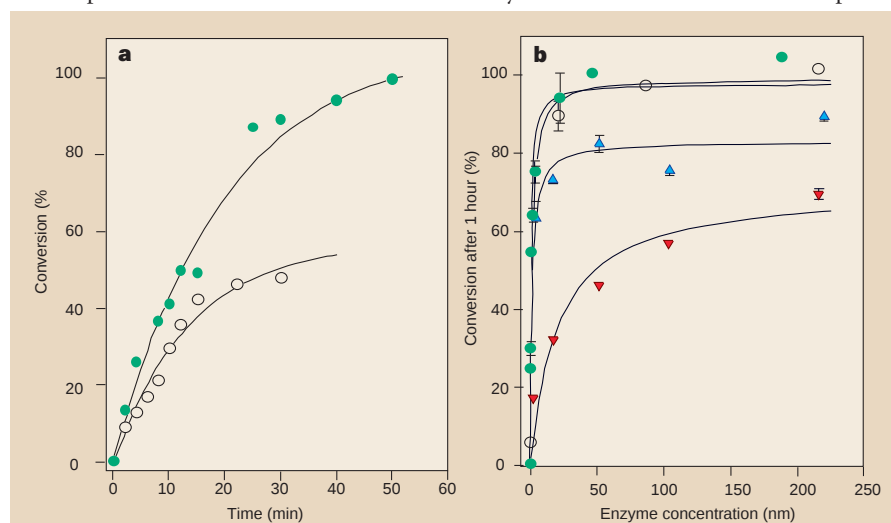


Figure 1 OPH foam performance on paraoxon-contaminated glass surfaces. **a**, Time dependence of surface decontamination. Initial paraoxon surface concentrations were 0.27 $\mu\text{mol cm}^{-2}$ (green circles) and 0.52 $\mu\text{mol cm}^{-2}$ (white circles). Foam was applied at a height of 1.2 cm. **b**, Conversion is increased by increasing enzyme loading within the foam. Enzyme foams (1.2 cm in height) were used to detoxify surfaces with higher degrees of contamination. Initial contamination levels included 0.26 $\mu\text{mol cm}^{-2}$ (green circles), 0.32 $\mu\text{mol cm}^{-2}$ (blue triangles), 0.75 $\mu\text{mol cm}^{-2}$ (white circles) and 1.51 $\mu\text{mol cm}^{-2}$ (red triangles).

1. Matsumoto, M., Fu, Y.-X., Molina, H. & Chaplin, D. D. *Immunol. Rev.* **156**, 137–144 (1997).
2. Alimzhanov, M. B. *et al.* *Proc. Natl. Acad. Sci. USA* **94**, 9302–9307 (1997).
3. Koni, P. A. *et al.* *Immunity* **6**, 491–500 (1997).
4. Aggarwal, B. B. & Natarajan, K. *Eur. Cytokine Netw.* **7**, 93–124 (1996).
5. Ware, C. F., Van Arsdale, T. L., Crowe, P. D. & Browning, J. L.

on the efficiency of surface decontamination using OPH foams.

Foam of dilute enzyme concentration produces greater than 99% detoxification of a mildly contaminated surface ($0.27 \mu\text{mol paraoxon cm}^{-2}$) in one hour (Fig. 1a). Such dilute enzyme concentrations are inadequate for increased levels of contamination when the foam is applied to a height of 1.2 cm. However, a higher degree of contamination was effectively treated with similar foams of increasing enzyme content. The attainable conversion in one hour under varying conditions is illustrated in Fig. 1b. Figure 1 shows that nerve-agent surface concentrations in the range expected from a nerve-agent attack ($0.3 \mu\text{mol cm}^{-2}$) would be satisfactorily decontaminated over a wide range of enzyme concentrations.

Does adding more enzyme foam to a fixed surface area accommodate a higher concentration of neurotoxins or increase the decontamination rate? If the reaction takes place only at the foam-surface interface, adding more foam would be of little use. However, if the foam and/or the aqueous fallen film extract paraoxon from the surface, and the reaction occurs throughout the foam, adding more foam should increase the decontamination rate. Increased application of foam not only provides greater release rates for enzyme and water from the foam, but also increases the rate of detoxification. A surface contaminated with $1.15 \mu\text{mol paraoxon cm}^{-2}$ (nearly four times the typical level of contamination) was detoxified by an OPH foam (11.4 nM) to a moderate degree ($43 \pm 5\%$ conversion) when a 1.2-cm foam height was used. Identical foam applied to a height of 3.0 cm achieved greater than 70% conversion in the same time. Because foam for fire-fighting applications and hazardous vapour containment are generally applied to heights exceeding 5 cm, our experiments are within typical operating parameters.

Our results show that OPH is catalytically active within fire-fighting foam. The primary limitation of the enzymatic bioremediation of chemical weapons lies in the strict specificity of a given enzyme for a particular substrate². This issue is being addressed by the production of foams with several enzymes of varying specificity. Such multicomponent enzyme foams will provide a safe, environmentally acceptable means of performing wide-area decontamination of nerve agents.

Keith E. LeJeune

Department of Chemical Engineering,
Carnegie Mellon University,
Pittsburgh, Pennsylvania 15213, USA

James R. Wild

Department of Biochemistry and Biophysics,
Texas A&M, College Station,
Texas 77843-2128, USA

Alan J. Russell

Department of Chemical and Petroleum
Engineering, and Center for Biotechnology and
Bioengineering, University of Pittsburgh,
Pittsburgh, Pennsylvania 15261, USA
e-mail: russell@civeng1.civ.pitt.edu

1. Alternative Technologies for the Destruction of Chemical Agents and Munitions (National Research Council Commission on Engineering and Technical Systems, Washington DC, 1993).
2. Orton, R. D. *Technology Report for Congress* (Department of the Army, Aberdeen Proving Ground, Maryland, 1994).
3. Chemical Weapons Destruction, Advantages and Disadvantages of Alternatives to Incineration (GAO/NSIAD-94-123, General Accounting Office, Washington DC, 1994).
4. Convention on the Prohibition of the Development, Production, Stockpiling, and the Use of Chemical Weapons and on their Destruction (United States Senate, Washington DC, 8th August 1994).
5. Grimsley, J. K., Rastogi, V. K. & Wild, J. R. *Bioremediation: Principles and Practice* pp. 577-614 (Technomic, Irvine, 1997).
6. Cheng, T. C. & Calomiris, J. J. *Enzyme Microb. Technol.* **18**, 597-601 (1996).
7. Hoskin, F. C. G., Waler, J. E., Dettbarn, W. D. & Wild, J. R. *Biochem. Pharmacol.* **49**, 711-715 (1995).
8. LeJeune, K. E. *et al.* *Biotechnol. Bioeng.* **54**, 105-114 (1997).
9. LeJeune, K. E. & Russell, A. J. *Biotechnol. Bioeng.* **51**, 450-457 (1996).
10. Xu, B., Wild, J. R. & Kenerley, C. M. J. *J. Ferment. Bioeng.* **81**, 473-481 (1996).
11. Lai, K., Stolorow, N. J. & Wild, J. R. *Arch. Biochem. Biophys.* **318**, 59-64 (1995).
12. Grimsley, J. K., Scholtz, J. M., Pace, C. N. & Wild, J. R. *Biochemistry* **36**, 14366-14374 (1997).

Nematode phylogeny and embryology

Blaxter *et al.*¹ presented the phylogenetic tree of the phylum Nematoda based on the sequences of small subunit (SSU) ribosomal DNA (see also ref. 2). At the level of higher-order classification, this tree is substantially different to traditional taxonomic systems³⁻⁵. The first branching of the SSU-based phylogenetic tree divides the nematodes into three big groups. Here we point out that three distinct patterns of early embryonic development reported for nematodes are in good agreement with this new classification. Differences in development could be more consistent with the natural nematode system than with the usual morphological criteria.

The three nematode groups resulting from the first branching of the SSU-based phylogenetic tree are: clade I (orders Dorylaimida, Mononchida, Mermithida and Trichocephalida); clade II (orders Enoplida and Triplonchida); and a group subdivided into clades III-V that consists of Chromadoria and Secernentea.

Development of nematodes from clade II is different from development of nematodes of all other groups because the divisions in the early embryos (up to the eight-cell stage) are synchronous and produce blastomeres indistinguishable from each other by size, position and appearance^{5,6}. Tracing cell fates with intracellular labels in group representatives

Enoplus brevis^{6,7} and *Pontonema vulgare* (our unpublished results) shows that the blastomeres at these stages have no regular cell-lineage pattern. One of the two first-formed blastomeres may contribute to anterior or posterior, left or right, or intermediate parts of the embryo, and the endoderm precursor segregated at the eight-cell stage may derive from either blastomere.

In contrast, nematode embryos from clades III-V and I generate blastomeres that are already clearly distinguishable from each other after the first division. Embryonic development of nematodes from clades III-V has been well studied, and, in *Caenorhabditis elegans*, the complete stereotyped cell lineage has been followed⁸. In nematode embryos from this group, the anterior blastomere is named AB; its progeny divides in a synchronized pattern and gives rise to the particular structures of the animal. Predominantly posterior and internal parts of the embryo, including endoderm, derive from another blastomere of the two-cell stage — P₁ (refs 5,8,9).

In nematodes from clade I, the first division also produces two blastomeres that are different to each other in size and fate, but are not homologous to AB and P₁ in nematodes of clades III-V. At the four-cell stage, two daughters of different blastomeres behave like AB progeny and another two resemble P₁ descendants. Endoderm in these species derives from the anterior blastomere of the first pair^{5,10,11}.

Thus, the cell-lineage patterns in the three large groups of nematodes seem to be very different. We realize that the above framework for different patterns of development in three taxonomic groups of nematodes is based on a few observations with varying degrees of reliability, but it gives straightforward, testable predictions.

D. A. Voronov, Yu. V. Panchin

Institute of Problems of Information Transmission,
Bolshoy Karetny per. 19, 101447 Moscow, Russia
e-mail: voronov@neuro.genebee.msu.ru

S. E. Spiridonov

Institute of Parasitology, Leninskii pr. 33,
117071 Moscow, Russia

1. Blaxter, M. L. *et al.* *Nature* **392**, 71-75 (1998).
2. Aleshin, V. V. *et al.* *Russian J. Nematol.* **6**, 175-184 (in the press).
3. Chitwood, B. G. & Chitwood, M. B. *Introduction to Nematology* (Univ. Park, Baltimore, 1974).
4. Lorenzen, S. *The Phylogenetic Systematics of Free-living Nematodes* (The Ray Society, London, 1994).
5. Malakhov, V. V. *Nematodes. Structure, Development, Classification, and Phylogeny* (Smithsonian Inst., Washington, 1994).
6. Voronov, D. A. & Panchin, Yu. V. *Development* **125**, 143-150 (1998).
7. Voronov, D. A. *et al.* *Proc. Acad. Sci. USSR* **286**, 201-204 (1986).
8. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. *Dev. Biol.* **100**, 64-119 (1983).
9. Skiba, F. & Schierenberg, E. *Dev. Biol.* **151**, 597-610 (1992).
10. Drozdovskii, E. M. *Proc. Acad. Sci. USSR* **222**, 1105-1108 (1975).
11. Malakhov, V. V. & Spiridonov, S. E. *Zool. Zhurnal* **60**, 1574-1577 (1981).

Sir Oran Haut-ton finds his voice

Apes, Language, and the Human Mind

by Sue Savage-Rumbaugh, Stuart G. Shanker and Talbot J. Taylor
Oxford University Press: 1998. 288 pp.
£22.50, \$29.95

Robert Seyfarth

In the first edition of the *Systema Naturae* (1735), Linnaeus arranged animals into groups based exclusively on anatomical features, lumping humans and nonhuman primates together. This sparked a strong reaction, particularly from the followers of René Descartes, who believed that humans were unique because they possessed minds and language. In later editions, Linnaeus responded by separating humans from other animals and basing his classification not just on anatomy but also on temperament, character, type of garments worn (if any), and forms of government.

In the nineteenth century, Cartesian dualism found an outlet in satirical accounts of apes who, if properly trained, could become respectable members of human society. In one story, an orang-utan named Sir Oran Haut-ton gains respect because he rescues a maiden in distress without taking advantage of her. Lacking language, however, he is not considered truly human. None the less, Sir Oran is elected to parliament, where his lack of speech becomes an asset because it gives him the reputation of a powerful but cautious thinker.

The authors of the present volume believe that Cartesian dualism is alive and well, particularly among those who refuse to accept Sue Savage-Rumbaugh's claim that Kanzi, a bonobo (or pygmy chimpanzee), has learned language. These critics, the authors claim, are wrong and their errors have major consequences: "When the implications of the Kanzi research have been fully assimilated, the way we look at, understand, and represent the relationships between language, cognition, and behavior will no longer be the same."

The first of the book's four chapters is a narrative account of Kanzi's infancy and youth. The second reviews Descartes's theories and argues that modern Cartesians — called 'bifurcationists' — dominate all branches of cognitive science. In chapter three, the authors describe what must be done "to place ape and human studies on an unbiased (that is, scientific) footing". In chapter four, they discuss the new perspectives that will emerge as a result.

Unlike subjects in earlier ape-language studies, Kanzi was not explicitly taught a linguistic form of communication. Instead, he 'picked up' the use of lexigrams on a

computer keyboard while Savage-Rumbaugh and her colleagues were trying (with some difficulty) to teach these lexigrams to Kanzi's mother. This is an important point because it means that Kanzi's ultimate performance cannot be dismissed as simply the result of training. In fact, more than any other captive ape, Kanzi seems to have learned his system of communication in ways that are similar to those used by children. Kanzi's communication is also unusual because he interacts with people in at least three ways: by using his keyboard, through hand gestures such as pointing, and by responding to what people say. Kanzi's linguistic skills are most strikingly displayed when he responds correctly to unusual commands such as "Put the toothbrush in the lemonade".

Many scientists, however, remain unconvinced. Not, the authors believe, because they have examined the data objectively, but instead because they share a pervasive Cartesian bias against the application of mentalistic terms to apes and the use of linguistic terms to describe their communication. Psychologists studying animal learning, for instance, accept that an animal's behaviour can be guided by mental representations like memories, but refuse to say that 'meaning' or 'reference' play any role in

the language of captive apes. This is inconsistent.

Similarly, the authors argue, the biased views of bifurcationist linguists allow them no room *in principle* for Savage-Rumbaugh's idea of 'primitive linguistic skills'. When she claims that Kanzi understands simple sentences, they reply that to understand a sentence one's brain must first parse it, and to parse it one must have an internalized generative grammar with all of its attendant parameters, rules, lexical categories, dictionary entries and so on. In other words, to prove that Kanzi has a partial language one must prove that he has it all; otherwise he doesn't have any.

To get around this methodological cul-de-sac, the authors advocate an entirely different approach to Savage-Rumbaugh's results. Instead of analysing each claim — that Kanzi understands action-object relations, for instance — separately, as a kind of 'propositional atom', results "must be seen... as emergent properties of the lived story that is partially recounted in [Savage-Rumbaugh's] narrative". If the discourse works, we should accept claims about its underlying elements. If Kanzi responds to a sentence the way a human would, and if a human's response to Kanzi's sentence doesn't shock or surprise him, then sentence comprehension



Sex and politics in the furred world

Screaming, hooting, scheming and arguments — no, it's not the House of Commons, but all part of the power struggles of the chimpanzee colony at Arnhem Zoo in The Netherlands. The revised edition of Frans de Waal's *Chimpanzee Politics* (John Hopkins University Press, £20.50)

expands and updates the story of the colony, now more than 25 years old, and reveals that the social gap between humans and chimpanzees is surprisingly small. Small wonder that US politician Newt Gingrich put the first edition on his reading list for members of Congress.

must have taken place. And (the authors quote Wittgenstein) "to understand a sentence means to understand a language".

Are the critics of ape-language research biased? I don't think so. In fact, the critical accounts written by Steven Pinker, Joel Wallman, Michael Tomasello and others have helped to elucidate the complexities of language and to clarify which linguistic elements are missing from Kanzi's impressive communication (for the most recent debate see E. Kako's forthcoming "Elements of syntax in the systems of three language-trained animals", *Anim. Learn. Behav.*, in the press).

Are cognitive scientists hopelessly wedded to Cartesian dualism? It is certainly true that practitioners in this new field do occasionally act as if humans are the only organisms with brains — one review, for example, *Foundations of Cognitive Science*, edited by M. I. Posner (MIT Press, 1989), contains no mention of animals at all — but this bias is changing. Most cognitive scientists now recognize that theories of how brains work must take account of ants that navigate across vast tracts of desert, birds that recover thousands of previously stored seeds, monkeys that recognize other animals' social relationships, as well as Kanzi.

Moreover, current views of animals' and children's ability to recognize the mental states of others are by no means as clear cut and Cartesian as the authors suggest. They make no mention, for example, of recent studies by Dare Baldwin, Simon Baron-Cohen, Daniel Povinelli, Tomasello and others that show that mental state attribution is not simply an ability that is either present or absent. Even in children, an awareness of others' mental states emerges gradually, progressing from early manifestations of social referencing by infants to the comprehension of false beliefs by four-year-olds.

Also not mentioned are studies by Tomasello showing that in general chimpanzees raised by humans perform cognitive tasks better than chimpanzees raised by chimpanzees, regardless of whether the animals have received language training. This reminds us that a complete understanding of the relation between language and cognition cannot in principle come from Kanzi or the ape-language projects alone.

Finally, there is an irony in the authors' view that full acceptance of their results is blocked by the Cartesian dualism of their critics. They demonstrate clearly that dualists come in many forms, but neglect to mention the group of Cartesians to which they themselves belong. For years, psychologists and linguists have assumed that the best way to explore an ape's mind and its capacity for language is to bring it into human society. In the nineteenth century, this view inspired imaginary, satirical accounts; in the twentieth, it drove the French to establish a model

village in French Guinea as a training ground to civilize apes (local women would be nurses and guides), the British to plan a colony to make chimpanzees human, and Americans to teach apes language. Few people would ever apply the same assumption to humans — we naturally study the minds and languages of people in other cultures in their own habitats and on their own terms. Viewed in this light, the ape-language studies of Savage-Rumbaugh and others are, in their methods if not their conclusions, classically Cartesian. From them we have learned an extraordinary amount about the cognitive and communicative skills of apes immersed in human society. What they tell us about the mind and communication of apes in ape society remains to be seen. □

Robert Seyfarth is in the Department of Psychology and the Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

Trials and errors

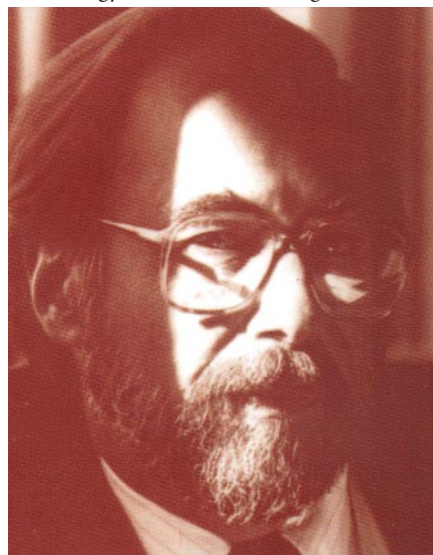
The Baltimore Case: A Trial of Politics, Science and Character

by Daniel J. Kevles

Norton: 1998. 509 pp. \$29.95, £21

Jon Turney

Few of us, I dare say, had the stamina to follow the Baltimore affair properly. A rich stew of supposedly fishy results, personal vendettas, enormous egos and the murky politics of leaks, smears and intimidation, it simmered for ten years. True, it had spice. An American Nobel laureate in biology was hauled before pugnacious Congressmen to rebut charges of scientific fraud in a paper he had co-authored. The US secret service pored over laboratory notebooks from no less an institution than the Massachusetts Institute of Technology (MIT). But the charges turned on



Baltimore: refused to be cowed by investigations based on shoddy evidence.

the interpretation of fairly arcane aspects of the expression of immunity in transgenic mice. And, by the time the multiple investigations, and investigations of the investigators, came to a halt, it was hard to feel any the wiser.

So Daniel Kevles performs a signal service by applying his historian's acumen to the mountain of documents accumulated by all those investigations, and to his own detailed interviews with the participants, to give us an orderly narrative of the whole affair. He did so, he says, because it seemed to him likely to "throw some light on science in late-twentieth-century American society".

But he clearly got caught up, as historians of contemporary affairs will, in the detail of the events, to the extent that he published a lengthy article in *The New Yorker* exonerating the accused scientists some weeks before a panel of the US National Institutes of Health came to the same conclusion. The result is a book that is likely to be judged, as one of the blurb writers suggests, the definitive account of the affair, but not, perhaps, a complete consideration of its wider significance.

So what was it all about? The first thing is that it should more properly be known as the Imanishi-Kari affair. David Baltimore's co-worker, Thereza Imanishi-Kari, was an expert in cellular immunology, and was accused of faking results published in *Cell* in a paper co-authored with Baltimore and four other biologists. In fact, the accusations grew stronger over time, but that was the final allegation. The results — and the scientists — were then subject to an extraordinary series of investigations: by Tufts University, Imanishi-Kari's new employers; by MIT; by the self-appointed 'fraud-busters' at the National Institutes of Health, Walter Stewart and Ned Feder; by a congressional committee; and by a succession of NIH panels and reviews.

The first two found no case to answer.



Imanishi-Kari: faced accusations from within her own laboratory.

The others convicted Imanishi-Kari and Baltimore variously of inaccuracy, scientific misjudgement, error, arrogance and fraud. After the results of a draft report by the NIH's newly fledged Office of Scientific Integrity were leaked in 1991, Baltimore was obliged to retract the paper (and, soon after, to resign his presidency of Rockefeller University in New York). After further thought, he retracted the retraction. Finally, an NIH appeals board reviewed all the charges — now numbering 19 — that had been laid against Imanishi-Kari by NIH investigators, and rejected every one.

Kevles is clear throughout that the *Cell* authors were victims of a miscarriage of justice, only put right because of their refusal to be cowed by weighty indictments based on shoddy evidence. He explains in detail how the final review showed that the NIH's supposedly exhaustive investigation had supplied no plausible motive for fraud. And its case rested almost entirely on forensic evidence (mostly from analysis of printer inks recording radiation counts and of the sequences of entries in lab notebooks) which itself failed to withstand serious scientific scrutiny. The data offered against Imanishi-Kari, in other words, fell below the standards her paper was being held to. Before recording the two biologists' vindication, though, Kevles has the more taxing task of explaining how the affair acquired such extraordinary momentum.

Imanishi-Kari's original accusers were from her own laboratory, her main antagonist being a postdoctoral researcher, Margaret O'Toole — later cast as a heroic whistleblower by almost all the reporters covering the story. Somehow, difficulty with some of the techniques used in the paper, construction of different interpretations of the results, and long contemplation of fragments of the data, coalesced to generate first suspicion, then certainty, that all was not as it seemed.

Kevles's account shows these suspicions being nurtured in an institutional hothouse, rife with thwarted ambition, personal animosity, jealousy, nepotism, double-dealing, secrecy and paranoia. He does not say whether these currents were stronger in this particular lab, at MIT, or are more or less those found in any corner of a modern biomedical research enterprise, where everyone is frantically chasing publications, tenure, grants, or just enough data for a PhD.

What he does show is how these suspicions were picked up and amplified by critics pressing the US scientific community to take fraud and misconduct more seriously, mistrustful of self-regulation, and determined that researchers must be accountable for all the NIH's billions.

Although he is critical of how most of these people conducted themselves, Kevles seems to feel that their reasons were mostly

valid. It was just that in this case they got the wrong people.

It is a sorry saga, the more so because Kevles is bound to tell us a story whose end we already know. But in hindsight it is not, I think, of enormous significance, except for the victims. The outcome helped to stem the over-zealous pursuit of scientific misconduct, but there were enough genuine cases to fuel continuing concern. The NIH's final verdict upheld the *Cell* authors' scientific right to publish what they judged to be the best interpretation of their results. But this did not overturn the broader arguments for researchers to submit to outside scrutiny.

What is the moral of the story? The most useful lesson is one that Kevles only hints at at the very end of the book. He quotes approvingly from Baltimore to the effect that deciding what counts as a worthwhile scientific paper is a matter of professional judgement, specific to a particular field at a particular time.

This fact of research life, it seems, was denied by the congressional inquirers, by many journalists, and by some of the accused workers' scientific peers. But, if all of these people appeared to believe that every research publication must contain certified and irrevocable Truth about how the world is, perhaps that is because that image of science has been found advantageous by most scientists most of the time.

If any good came of the unsavoury affair, it may be the extended public exposure of what scientific practice is actually like — full of ambiguous results, unexplained anomalies, imprecise assays, and of the craft skill which turns all these things into a story fit for journal referees. As Baltimore said, however a paper makes it look, researchers know that "no study is ever complete". □

Jon Turney is in the Department of Science and Technology Studies, University College London, Gower Street, London WC1E 6BT, UK.

A cold fish among the trilobites

**Charles Doolittle Walcott,
Paleontologist**

by Ellis L. Yochelson

Kent State University Press: 1998. 510 pp.

\$49, £43.95

Richard Fortey

Seldom has a middle name been less appropriate than in the case of Charles Doolittle Walcott. As a geologist, palaeontologist and politician, the man was indefatigable. He was the very model of self-improvement, even by nineteenth-century standards, when pulling oneself up by diligence and application was at its acme. From humble, if middle-class, beginnings — and without a university degree — Walcott progressed from assistant to the great but flawed James Hall (a man who combined the qualities of pioneer, scientist and mountebank in equal measure) to become, successively, director of the US Geological Survey and director of the National Museum of Natural History at the Smithsonian Institution in Washington, DC.

If ever there was an establishment figure in turn-of-the-century geology, Walcott was that man. Yet the same man carried out pioneering fieldwork on the geology of the Grand Canyon under conditions of considerable hardship, won tough intellectual battles over the recognition of the Cambrian and Ordovician systems in America, and described the limbs of trilobites for the first time. Even if he had done no more than publish scientific papers his place in the history of science would have been assured.

Today, he is known principally as the discoverer of the famous Burgess Shale, in British Columbia, where the fossils of soft-bodied Cambrian animals are miraculously



Beyond the Shale: Walcott's place in history was assured even had he not discovered the Burgess Shale.

preserved. Walcott himself was so insecure about academic acceptance of the reality of these forms that he sent specimens to famous contemporary collections around the world to prove that no fakery was involved. Such circumspect behaviour was typical of him.

In this biography, Ellis Yochelson plots Walcott's prior career in painstaking detail — a career that embraced administration as much as research, political lobbying as much as geological theorizing. Yochelson paints an interesting and minutely observed picture of geological science on the hoof in its heroic days. It may not seem so unusual to us Thatcherite survivors today, but Walcott was adept at selling the economic benefits of studies of Cambrian brachiopods, or Chinese trilobites, in ways that would leave our scientific bureaucrats dumbfounded. He lobbied senators and academics alike, progressing daily from gruelling committee hearings to suppers at exclusive Washington clubs, pursuing the cause of the Geological Survey by diplomacy and example.

Yochelson's account of the political shenanigans of the time is admirable — if a little wearing. Given Walcott's relentless daily shuffling among Washington's great and good, it occasionally tires the reader to be attached so closely to his coat tails. But such careful attention does illuminate the dedication of the true administrator. These days, we may be accustomed to think of the chief as the enemy of creativity, the arbiter who stifles imagination. No such distinction was apparent to Walcott, whose every administrative triumph seemed to be followed by a scientific coup.

Unfairly, perhaps, I longed for Yochelson to reveal Walcott's Achilles' heel. Was he a secret womanizer? Did he have an uncontrollable temper, perhaps? A weakness for rye? Apparently, he was a paragon of virtue in all departments. He won his intellectual battles by means of reason and astute alliances, he mastered complex political briefs, and in the interstices composed magisterial monographs; he loved his family loyally and deeply. His diary entries are mostly those of a cold fish — he never utters a hurrah even when awarded honorary doctorates or prestigious medals. The entries warm to a little more than tepid only when he records his grief over the loss of his first wife.

The real disappointment of Yochelson's life of Walcott is that it is only about 70 per cent of a biography. It stops suddenly at Walcott's appointment to the Smithsonian Institution. This is a pity — not least because we do not reach the later science, including the Burgess Shale. Maybe a second volume is promised? The scholarship is admirable, at least from the North American side. It is disappointing to find the 'British Walcott', Archibald Geikie, persistently misspelled, and the spelling of Welsh placenames is very

approximate indeed. But Walcott's is a life that needed to be described, and Yochelson has more than done him justice. □

Richard Fortey is in the Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

Animal analysts who know their plays

Game Theory and Animal Behaviour

edited by Lee Alan Dugatkin and Hudson Kern Reeve
Oxford University Press: 1998. 320 pp. £55, \$78.

Evolutionary Games and Population Dynamics

by Josef Hofbauer and Karl Sigmund
Cambridge University Press: 1998. 323 pp.
£50, \$69.95 (hbk); £16.95, \$27.95 (pbk)

Marc Mangel

Mathematical models of all sorts have contributed deeply to our understanding of biology. These two volumes deal with game theory, which is the branch of optimization models (formalized by von Neuman and Morgenstern in 1944) in which the behaviour of an individual depends on the frequency of certain actions of other individuals in the population. Although it was originally envisioned as an economic theory, John Maynard Smith recognized in the early 1970s that formal game theory had much to offer to biology, particularly because some of the concepts that had only vague definition in economics, such as utility, had very clear definitions and implications in biology, such as fitness. The ideas of Maynard Smith, based on the notion of an evolutionarily stable strategy (roughly, a strategy that, when adopted by all the members of a population, prevents the invasion through natural selection of any alternative strategy), form the foundation of evolutionary game theory.

Mathematical applications in biology are as diverse as biology itself, ranging from knot theory through partial differential equations to optimization methods. Donald Ludwig once wrote, "the misleading term 'Mathematical Biology' suggests that some part of Biology is being taken over by Mathematics. Biological Mathematics would be a better term. Linear thinking and methods still dominate the applications of mathematics. Biological requirements deflect us from that easy path". Both books leave the easy path behind.

Dugatkin and Reeve's edited collection explores various mathematical methods and models in evolutionary game theory, including the Prisoner's Dilemma, the Hawk-Dove game, the Sequential Assessment game, and the War of Attrition Model. But it also describes many interesting examples of real-

life animal behaviour, including games between foraging producers and scroungers, reciprocal grooming in impala, territorial defence by birds and spiders, animal communication, parent-offspring conflict, and colony founding by ants. There are many accounts of experimental tests of game theory models, along with clear discussions of the limitations of the game theory approach.

The quality of writing (often a problem in edited volumes) is uniformly good. The chapter by R. Gomulkiewicz is especially important, because it connects game theory, other optimization methods, and quantitative genetics with a focus on an empirical strategy for detecting adaptation and constraint. These different mathematical methods are often described in opposition to each other, when they should be viewed as different ways of seeing the same problem.

The starting point of Hofbauer and Sigmund's book is the replicator equation, which recognizes that the rate of change of a certain type in a population depends on its frequency and its expected reproduction, relative to the other types in the population. Many mathematical results from the past decade are included, so the book is up to date and could be used for a course in mathematical aspects of theoretical ecology. The book has two introductions: one for game theorists and one for biologists. Even so, most biologists will find it hard going. The first part could be used in an advanced course on population dynamics or ecological modelling; but even here there is some sophisticated mathematics and heavy-duty notation. For example, the fourth chapter begins with the Poincaré-Bendixson theorem and continues to Hopf's bifurcation theorem; these are elegant, important and beautiful theorems, but difficult.

In the middle two sections of the book, the authors provide an elegant tour of evolutionary game theory and its connections to the classical equations of population dynamics. They provide a few specific examples, such as the rock-scissors-paper game and the repeated Prisoner's Dilemma, which may help bring the mathematics into clearer focus.

The last part of the book connects dynamical systems and frequency dependent genetics, playing much the same role as the article by Gomulkiewicz in showing the connections between dynamical systems and genetics.

Reading these two books together is interesting because it focuses the difference between theoretical biology (Dugatkin and Reeve) and mathematical biology (Hofbauer and Sigmund) in much the same way that theoretical and mathematical physics differ. Together, they have something for anyone interested in game models in organismal biology. □

Marc Mangel is in the Department of Environmental Studies, University of California, Santa Cruz, CA 95064, USA.

Origin of quantum-mechanical complementarity probed by a 'which-way' experiment in an atom interferometer

S. Dürr, T. Nonn & G. Rempe

Fakultät für Physik, Universität Konstanz, 78457 Konstanz, Germany

The principle of complementarity refers to the ability of quantum-mechanical entities to behave as particles or waves under different experimental conditions. For example, in the famous double-slit experiment, a single electron can apparently pass through both apertures simultaneously, forming an interference pattern. But if a 'which-way' detector is employed to determine the particle's path, the interference pattern is destroyed. This is usually explained in terms of Heisenberg's uncertainty principle, in which the acquisition of spatial information increases the uncertainty in the particle's momentum, thus destroying the interference. Here we report a which-way experiment in an atom interferometer in which the 'back action' of path detection on the atom's momentum is too small to explain the disappearance of the interference pattern. We attribute it instead to correlations between the which-way detector and the atomic motion, rather than to the uncertainty principle.

In classical physics, a particle moves along a well-defined trajectory. A quantum object, however, reveals its wave character in interference experiments in which the object seems to move from one place to another along several different paths simultaneously. It is essential that these ways are indistinguishable, because any attempt to observe which way the object actually took unavoidably destroys the interference pattern.

The usual explanation for the loss of interference in a which-way experiment is based on Heisenberg's position-momentum uncertainty relation. This has been illustrated in famous *gedanken* experiments like Einstein's recoiling slit¹ or Feynman's light microscope². In the light microscope, electrons are illuminated with light immediately after they have passed through a double slit with slit separation d . A scattered photon localizes the electron with a position uncertainty of the order of the light wavelength, $\Delta z \approx \lambda_{\text{light}}$. Owing to Heisenberg's position-momentum uncertainty relation, this localization must produce a momentum uncertainty of the order of $\Delta p_z \approx h/\lambda_{\text{light}}$. This momentum uncertainty arises from the momentum kick transferred by the scattered photon. For $\lambda_{\text{light}} < d$, which-way information is obtained, but the momentum kick is so large that it completely washes out the spatial interference pattern.

However, Scully *et al.*³ have recently proposed a new *gedanken* experiment, where the loss of the interference pattern in an atomic beam is not related to Heisenberg's position-momentum uncertainty relation. Instead, the correlations between the which-way detector and the atomic beams are responsible for the loss of interference fringes.

Such correlations had already been studied experimentally. They are, for example, responsible for the lack of ground-state quantum beats in time-resolved fluorescence spectroscopy⁴. Other examples are neutron interferometers, where which-way information can be stored by selectively flipping the neutron spin in one arm of the interferometer^{5,6}.

Nevertheless, the *gedanken* experiment of Scully *et al.* was criticized by Storey *et al.*⁷, who argued that the uncertainty relation always enforces recoil kicks sufficient to wash out the fringes. This started a controversial discussion^{8–11} about the following question: "Is complementarity more fundamental than the uncertainty

principle?"¹⁰. This motivated Wiseman *et al.*^{10,11} to investigate what constitutes a momentum transfer in a double-slit experiment. They call the usual momentum transfer, like that in Feynman's light microscope, a "classical" kick; in addition, they define the concept of "quantum" momentum transfer. They find that the loss of interference need not be due to "classical" kicks. In this case the "quantum" momentum transfer cannot be less than that required by the uncertainty principle, so that these "quantum" kicks wash out the fringes.

In this context, Eichmann *et al.*¹² performed an experiment with a light interferometer, where the double slit is replaced by two trapped ions, which can store which-way information in internal states. This scheme was criticized⁹, because the ions play a double role: they act as sources of elementary waves (just like a double slit) and simultaneously as a which-way detector. Hence the momentum transfer from the double slit and the which-way detector cannot be separated.

Here we report on a which-way experiment with an atom interferometer. A microwave field is used to store the which-way information in internal atomic states. We study the mechanical effect of the which-way detection on the atomic centre-of-mass motion separately, and find that the "classical" momentum kicks are much too small to wash out the interference pattern. Instead, correlations between the which-way detector and the atomic motion destroy the interference fringes. We show that the back action onto the atomic momentum implied by Heisenberg's position-momentum uncertainty relation cannot explain the loss of interference.

The atom interferometer

Figure 1 shows a scheme of our atom interferometer. An incoming beam of atoms passes through two separated standing wave light beams. The detuning of the light frequency from the atomic resonance, $\Delta = \omega_{\text{light}} - \omega_{\text{atom}}$, is large so that spontaneous emission can be neglected. The light fields each create a conservative potential U for the atoms, the so-called light shift, with $U \propto I/\Delta$, where I is the light intensity (see, for example, ref. 13). In a standing wave the light intensity is a function of position, $I(z) = I_0 \cos^2(k_{\text{light}}z)$, where k_{light} is the wavevector of the light. Hence the light shift potential takes the form $U(z) = U_0 \cos^2(k_{\text{light}}z)$, with $U_0 \propto I_0/\Delta$.

The atoms are Bragg-reflected from this periodic potential, if they enter the standing light wave at a Bragg angle (see, for example, ref. 14). This process is similar to Bragg reflection of X-rays from the periodic structure of a solid-state crystal, but with the role of matter and light exchanged. In our experiment, the light creates the periodic structure, from which the matter wave is reflected.

The standing light wave splits the incoming atomic beam A (see Fig. 1) into two beams, a transmitted beam C and a first-order Bragg-reflected beam B. The angle between the beams B and C corresponds to a momentum transfer of exactly $2\hbar k_{\text{light}}$, as determined by the spatial period of $U(z)$. By varying the light intensity, the fraction of reflected atoms can be adjusted to any arbitrary value. In our experiment, the reflectivity of the beam splitter is tuned to $\sim 50\%$.

After switching off the first standing light wave, the two beams are allowed to propagate freely for a time interval t_{sep} . During this time, beam B moves a horizontal distance $d/2$ to the left and beam C moves $d/2$ to the right. The longitudinal velocities (vertical in Fig. 1) of the two beams are not affected by the light field. Then a second standing light wave is switched on, which also serves as a 50% beam splitter. Now two atomic beams D and E are travelling to the left, while beams F and G are travelling to the right. In the far field, each pair of overlapping beams produces a spatial interference pattern. The fringe period is the same as in a double-slit experiment with slit separation d . The relevant wavelength is the de Broglie wavelength associated with the momentum of the atoms. The envelope of the fringe pattern is given by the collimation properties of the initial atomic beam A.

The experiment is performed with the apparatus described in ref. 15: ^{85}Rb atoms are loaded into a magneto-optical trap (MOT). After trapping and cooling, the cloud of atoms is released and falls freely through the apparatus. The resulting pulsed atomic beam is collimated with a mechanical slit 20 cm below the MOT. The atoms then pass the interaction region with the standing light wave inside a microwave resonator. In the far field of the interaction region, 45 cm below the MOT, the atomic position distribution is observed by exciting the atoms with a resonant laser beam and detecting the fluorescence photons. The interaction time t_{Bragg} of the atoms with the standing light wave is controlled by switching the light on and off. The small atomic velocity of 2 m s^{-1} in the

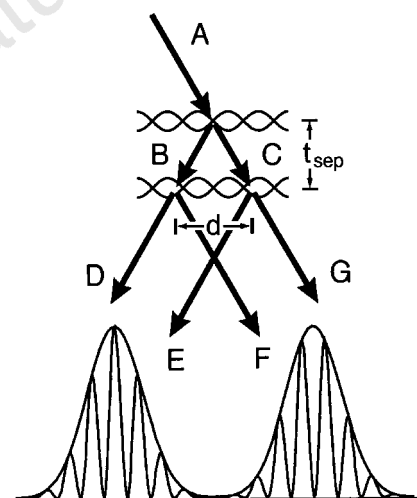


Figure 1 Scheme of the atom interferometer. The incoming atomic beam A is split into two beams: beam C is transmitted and beam B is Bragg-reflected from a standing light wave. The beams are not exactly vertical, because a Bragg condition must be fulfilled. After free propagation for a time t_{sep} the beams are displaced by a distance d . Then the beams are split again with a second standing light wave. In the far field, a spatial interference pattern is observed.

interaction region allows us to perform the whole interferometer experiment with only one standing light wave, which is switched on and off twice. As compared to ref. 15, only a few changes have been made. The width of the collimation slit was enlarged to $450\text{ }\mu\text{m}$. In order to improve the position resolution, the horizontal waist of the detection laser beam was reduced to $w = 50\text{ }\mu\text{m}$, and a second collimation slit with a width of $100\text{ }\mu\text{m}$ was added 1 cm below the MOT. Finally, the preparation of the internal atomic state was improved by removing atoms in wrong Zeeman sublevels.

Figure 2 shows the spatial fringe pattern in the far field for two different values of t_{sep} . We note that the observed far-field position distribution is a picture of the atomic transverse momentum distribution after the interaction.

Storing which-way information

A second quantum system is now added to the interferometer in order to store the information whether the atom moved along way B

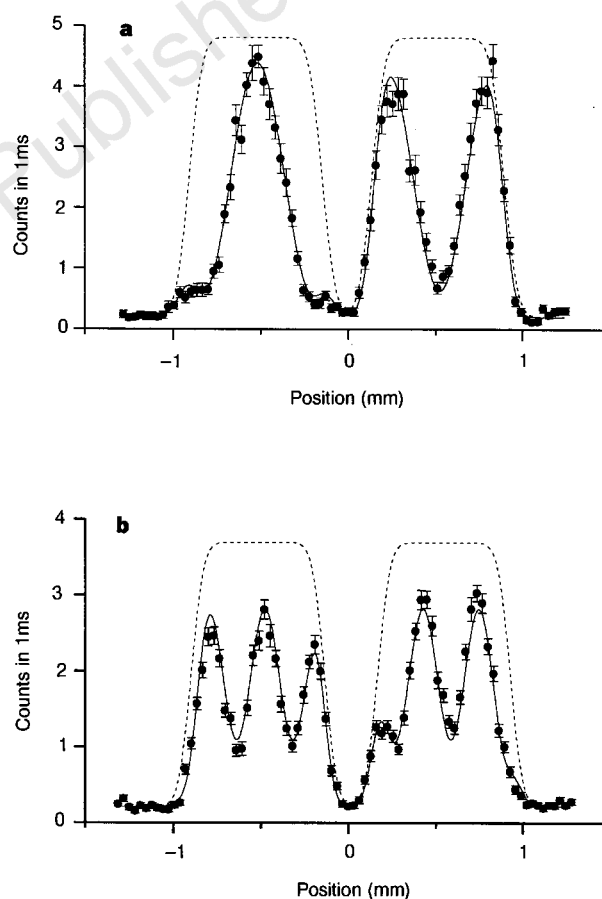


Figure 2 Spatial fringe pattern in the far field of the interferometer. The data were obtained with $t_{\text{Bragg}} = 45\text{ }\mu\text{s}$, and the 50% beam splitter was realized with $|U_0| = \hbar\pi/t_{\text{Bragg}}$. We chose $t_{\text{sep}} = 105\text{ }\mu\text{s}$ with $d = 1.3\text{ }\mu\text{m}$ (a), and $t_{\text{sep}} = 255\text{ }\mu\text{s}$ with $d = 3.1\text{ }\mu\text{m}$ (b). In both cases, the fringe period is in good agreement with the theoretical expectation. Each solid line represents a fit to the experimental data. The best-fit values for the visibilities are $(75 \pm 1)\%$ and $(44 \pm 1)\%$, respectively. The reduced visibility for the case of the narrow fringes is due to the finite position resolution of our apparatus. The dashed lines represent the independently measured beam envelope, which consists of two broad peaks. The right peak is due to beams F and G (see Fig. 1), with a shape determined by the momentum distribution of the initial beam A. The left peak is a combination of beams D and E. It is a Bragg-reflected picture of the right peak. The fringe patterns under these two broad peaks are complementary, that is, the interference maxima in the left peak correspond to interference minima in the right peak, and vice versa.

or C. Two internal electronic states of the atom are used as a which-way detector system. A simplified level scheme of ^{85}Rb is shown in Fig. 3a. Rabi oscillations between states $|2\rangle$ and $|3\rangle$ can be induced by applying a microwave field at ~ 3 GHz. To describe the information storing process, we first investigate the properties of one single Bragg beam splitter, using a simple model, whose validity will be discussed later.

The frequency of the standing light wave, ω_{light} , is tuned halfway between the $|2\rangle \rightarrow |e\rangle$ and $|3\rangle \rightarrow |e\rangle$ transitions. Hence the detunings from these transitions, Δ_{2e} and Δ_{3e} , have the same absolute value but opposite sign. The reflectivity of the beam splitter, that is, the probability of reflecting an atom, depends on $t_{\text{Bragg}} \times |U_0|$, and it is independent of the internal state.

However, the amplitude of the wavefunction experiences a phase shift which depends on the internal atomic state. A simple analogy for this phase shift can be found in light optics: a light wave reflected from an optically thicker medium experiences a phase shift of π , while reflection from an optically thinner medium or transmission into an arbitrary medium does not cause any phase shift. This argument also applies in atom optics: in our experiment, an atom in $|2\rangle$ sees a negative light shift potential (because $\Delta_{2e} < 0$), corresponding to an optically thicker medium, while an atom in $|3\rangle$ sees a positive potential (because $\Delta_{3e} < 0$), corresponding to an optically thinner medium. Hence an atom will experience a π phase shift only if it is reflected in $|2\rangle$.

This phase shift can be converted into a population difference between the hyperfine levels. For that purpose two microwave $\pi/2$ pulses resonant with the hyperfine transition are applied. They form a Ramsey scheme as shown in Fig. 3b. The atom is initially prepared in state $|2\rangle$. Then a microwave $\pi/2$ pulse is applied, converting the internal state to the superposition state $(|3\rangle + |2\rangle)/\sqrt{2}$. After this, the atom interacts with the standing light wave. As explained above, the atom will experience a π phase shift only if it is reflected and in state $|2\rangle$. Thus the internal state of the reflected beam is changed to $(|3\rangle + |2\rangle)/\sqrt{2}$, while the internal state of the transmitted beam is not affected. As a result, there is an entanglement created between the internal and the external degree of freedom of the atom. The state vector of the system becomes:

$$|\psi\rangle \propto |\psi_B\rangle \otimes (|3\rangle - |2\rangle) + |\psi_C\rangle \otimes (|3\rangle + |2\rangle) \quad (1)$$

where $|\psi_B\rangle$ and $|\psi_C\rangle$ describe the centre-of-mass motion for the reflected and transmitted beams (see Fig. 1), respectively. This entanglement is the crucial point for the storage of information. The second microwave pulse acting on both beams (the transmitted and the reflected), converts the internal state of the transmitted beam to state $|3\rangle$, while the reflected beam is converted to state

$|2\rangle$. Thus, the state vector after the pulse sequence shown in Fig. 3b becomes:

$$|\psi\rangle \propto -|\psi_B\rangle \otimes |2\rangle + |\psi_C\rangle \otimes |3\rangle \quad (2)$$

We note that momentum transfer from the microwave slightly changes $|\psi_B\rangle$ and $|\psi_C\rangle$, but this has negligible effects, as will be discussed below.

Equation (2) shows that the internal state is correlated with the way taken by the atom. The which-way information can be read out later by performing a measurement of the internal atomic state. The result of this measurement reveals which way the atom took: if the internal state is found to be $|2\rangle$, the atom moved along beam B, otherwise it moved along beam C.

A detailed calculation of the beam splitter reveals an additional phase shift, not discussed so far. It arises because the atoms travel in the light shift potential during t_{Bragg} . This creates a phase shift for the atoms proportional to $U_0 \times t_{\text{Bragg}}$, resulting in a relative phase shift between atoms in states $|2\rangle$ and $|3\rangle$. Fortunately, this phase shift is identical for the transmitted and reflected beams and therefore does not affect the storing process in an essential manner. Moreover, a small detuning of the microwave frequency from the atomic resonance allows us to compensate for this effect, so that the simple model discussed above is valid.

Interferometer with which-way information

After considering a single beam splitter, we now return to the complete interferometer. Sandwiching the first Bragg beam splitter between two microwave $\pi/2$ pulses stores the which-way information in the internal atomic state, as described above. We note that the second Bragg beam splitter does not change the internal state.

Will there still be interference fringes in the far field, when the which-way information is stored? The experimental result is shown in Fig. 4: there are no fringes. The data were recorded with the same parameters as in Fig. 2a. The only difference is that two microwave pulses were added to store the which-way information. Atoms in both hyperfine states were detected, so that the which-way detector was not read out. The mere fact that which-way information is stored in the detector and *could* be read out already destroys the interference pattern. We have verified experimentally that the fringes also disappear when the which-way detector is read out, that is, when only atoms in state $|2\rangle$ or only atoms in state $|3\rangle$ are detected. Of course, the absolute size of the signal is reduced by a factor of two in these cases.

Mechanical effects

We now discuss whether the loss of interference can be explained by mechanical effects of the which-way detector on the atomic

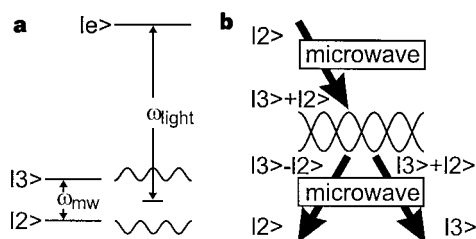


Figure 3 Storage of which-way information. **a**, Left, simplified level scheme of ^{85}Rb . The excited state ($5^2P_{3/2}$) is labelled $|e\rangle$. The ground state ($5^2S_{1/2}$) is split into two hyperfine states with total angular momentum $F = 2$ and $F = 3$, which are labelled $|2\rangle$ and $|3\rangle$, respectively. Right, the standing light wave with angular frequency ω_{light} induces a light shift for both ground states which is drawn as a function of position. **b**, The beam splitter produces a phase shift that depends on the internal and external degree of freedom. A Ramsey scheme, consisting of two microwave $\pi/2$ pulses, converts this phase shift into a population difference (see text).

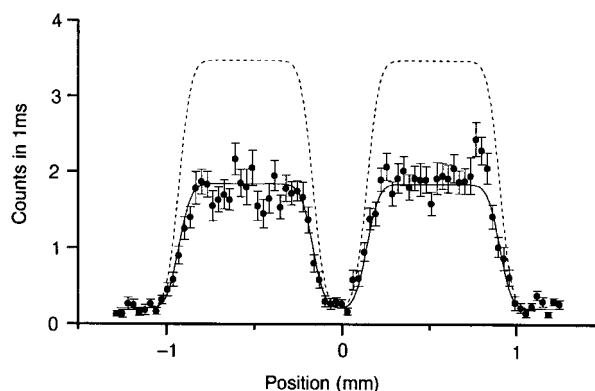


Figure 4 Same as Fig. 2a, but with which-way information stored in the internal atomic state. The interference fringes are lost due to the storage of which-way information.

centre-of-mass motion. Therefore the transverse momentum transfer in the microwave field and the longitudinal displacement of the atomic wavefunction must be investigated.

To wash out the interference fringes, the transverse momentum transfer must have a distribution whose spread must correspond to at least half a fringe period. Such a “classical” momentum transfer distribution would also broaden the envelope of the fringe pattern by the same amount¹¹. Comparing the experimental data in Fig. 2a and Fig. 4, it is obvious that the width of the envelope is not changed. This experimental result clearly shows that there is no significant transfer of transverse momentum in the microwave field.

The momentum transfer during the interaction with the microwave field can also be estimated theoretically. The microwave field is a standing wave diffracting the atomic beam. Because the atoms are travelling much less than a microwave wavelength during the interaction, the Raman–Nath approximation is valid. For a plane atomic wave, the probability to pick up n photon momenta during a $\pi/2$ pulse is $J_n^2(\pi/2)$, where J_n is the n th order Bessel function¹⁶. Hence the probability of transferring more than two microwave photon momenta is less than 1%. The absorption of a single microwave photon shifts the position of the atom in the detector plane by 5 nm. It follows that the mechanical recoils during the interaction with the microwave field can shift the pattern by at most ± 10 nm—too small to be observed. Of course, the transverse width of the atomic beam is much less than a microwave wavelength, so that the atomic beam is not a plane wave. However, the atomic beam is a superposition of plane waves. The above Raman–Nath calculation applies for each plane wave component, so that the spread of the atomic beam cannot be larger than for a plane wave.

The second mechanical effect that could explain the loss of interference is a longitudinal displacement of the atomic wavefunction, as discussed in ref. 7. In our experiment, this argument does not apply, because the whole interaction sequence is pulsed. The atom’s potential energy varies as a function of time, not as a function of longitudinal position, so that the interaction does not create any longitudinal forces and displacements.

We conclude that the “classical” mechanical effects of the which-way detector on the atomic centre-of-mass motion are negligible, so that some other mechanism must enforce the loss of interference fringes.

Correlations destroy interference

In order to investigate why the interference is lost, we consider the state vector for the interaction sequence used in Fig. 4. The state vector after the interaction with the first beam splitter sandwiched between the two microwave pulses is given in equation (2). The second beam splitter transforms this state vector into:

$$|\psi\rangle \propto -|\psi_D\rangle \otimes |2\rangle + |\psi_E\rangle \otimes |3\rangle + |\psi_F\rangle \otimes |2\rangle + |\psi_G\rangle \otimes |3\rangle \quad (3)$$

The sign of $|\psi_F\rangle$ is positive due to the π phase shift during the reflection from the second beam splitter.

In the far field, the atomic position distribution under the left peak of the envelope is given by:

$$P(z) \propto |\psi_D(z)|^2 + |\psi_E(z)|^2 - \psi_D^*(z)\psi_E(z)\langle 2|3\rangle - \psi_E^*(z)\psi_D(z)\langle 3|2\rangle \quad (4)$$

because here the spatial wavefunctions $\psi_F(z)$ and $\psi_G(z)$ vanish. The first two terms describe the mean intensity under the envelope. Interference could only be created by the last two terms, but they vanish because $\langle 2|3\rangle = 0$. Precisely the same entanglement that was required to store the which-way information is now responsible for the loss of interference. In other words: the correlations between the which-way detector and the atomic motion destroy the interference, as discussed in ref. 3.

This loss of interference manifests itself as a dramatic change in the momentum distribution when adding the microwave fields to the interferometer, even though the microwave itself does not transfer enough momentum to the atom to wash out the fringes.

However, the addition of the microwave fields modifies the probability for momentum transfer by the light fields. This modification of the momentum transfer probability is due to the correlations between the which-way detector and the atomic motion.

Correlations between the interfering particle and the detector system are produced in any which-way scheme, for example, in the previously mentioned *gedanken* experiments of Einstein’s recoiling slit and Feynman’s light microscope. But in these experiments “classical” mechanical effects of the detector on the particle’s motion can explain the loss of interference as well, so that the effect of the correlations is hidden.

So far, the microwave field has been treated as a classical field, that is, not as a quantized field. At first glance this might seem to be unjustified, because the population difference between the two hyperfine states corresponds to the absorption of one microwave photon, so that the system also becomes entangled with the microwave field. Hence the interference terms in equation (4) must include an additional factor $\langle \alpha|\beta\rangle$, where $|\alpha\rangle$ denotes the initial state of the microwave field, which changes to $|\beta\rangle$ due to the absorption of one photon. In our experiment, the initial state $|\alpha\rangle$ is a coherent state with a large mean photon number, and therefore the spread of the photon number is also large. It follows that $\langle \alpha|\beta\rangle \approx 1$, so that the entanglement with the microwave field has negligible effects, as has already been pointed out in ref. 3.

Uncertainty relation

We now discuss the role of Heisenberg’s position–momentum uncertainty relation in our experiment. For this discussion it is essential that the two ways through the interferometer (beams B and C in Fig. 1) are never separated in transverse position space. This is because the beams have a transverse width of 450 μm (as determined by the width of the lower collimation slit), but are shifted transversely only by a few micrometres (given by d) with respect to each other. This has an important consequence: the storage of which-way information does not imply any storage of transverse position information.

Moreover, the atom is not localized with a precision of the order of d at any stage of the experiment. In particular, the atom stays delocalized during its whole passage through the interaction region, regardless of whether the microwave is on or off. Heisenberg’s position–momentum uncertainty relation could only be invoked if the atom were localized. Hence the uncertainty relation does not imply any back action onto the transverse momentum, either “classical” or “quantum”.

Of course, in every real experiment there are localization effects due to the finite size of the apparatus. This localization leads to a back action onto the momentum. But this back action is not at all related to the fringe separation. For example, in our set-up the atoms are localized within one wavelength of the microwave. The corresponding back action onto the transverse momentum implied by the uncertainty relation is of the order of one microwave photon recoil and has been analysed above. This back action is four orders of magnitude smaller than the fringe separation, so that it cannot explain the loss of interference. Hence correlations can explain the loss of interference, while the uncertainty relation cannot. This answers the controversial question cited in the introduction: complementarity is not enforced by the uncertainty relation.

This result is not in conflict with the results of Wiseman *et al.*¹¹, who considered only “experiments in which double-slit interference patterns are destroyed by making a position measurement”¹¹. In our experiment, no double slit is used and no position measurement is performed, so that the results of ref. 11 do not apply. It is an open question whether the concept of “quantum” momentum transfer can be generalized to schemes without a mechanical double slit. Such a generalization would have to take into account the fact that in our experiment, the amount of momentum transferred by the light fields is always either zero or exactly $2\hbar k_{\text{light}}$.

Although beams B and C are never separated in transverse position space, they are separated in transverse momentum space. The separation is $2\hbar k_{\text{light}}$, as is required for first-order Bragg reflection. Storing which-way information therefore corresponds to storing transverse momentum information with an accuracy of the order of $\Delta p_z \approx \hbar k_{\text{light}}$. So the uncertainty relation implies that the storing process must include a back action onto the transverse position of the order of $\Delta z \approx \lambda_{\text{light}}$. This back action is due to the following effect. In the Bragg regime, the interaction time with the standing light wave, t_{Bragg} , is so long that the atoms move at least a transverse distance of the order of $\lambda_{\text{light}}/2$ within t_{Bragg} . In a naive picture, the atoms can be Bragg-reflected at the beginning or at the end of this interaction, which implies a transverse position uncertainty of the order of $\Delta z \approx \lambda_{\text{light}}$. But this back action onto the near-field position cannot destroy the far-field fringe pattern. The interference pattern created in the interferometer is a pattern in momentum space, not in position space. The far-field position distribution is simply a picture of the final momentum distribution. □

Received 3 February; accepted 15 June 1998.

1. Bohr, N. in *Albert Einstein: Philosopher-Scientist* (ed. Schilpp, P. A.) 200–241 (Library of Living Philosophers, Evanston, 1949); reprinted in *Quantum Theory and Measurement* (eds Wheeler, J. A. & Zurek, W. H.) 9–49 (Princeton Univ. Press, 1983).
2. Feynman, R., Leighton, R. & Sands, M. in *The Feynman Lectures on Physics* Vol. III, Ch. I (Addison Wesley, Reading, 1965).

3. Scully, M. O., Englert, B. G. & Walther, H. Quantum optical tests of complementarity. *Nature* **351**, 111–116 (1991).
4. Haroche, S. in *High-resolution Laser Spectroscopy* (ed. Shimoda, K.) 253–313 (Topics in Applied Physics, Vol. 13, Springer, 1976).
5. Rauch, H. et al. Verification of coherent spinor rotation of fermions. *Phys. Lett. A* **54**, 425–427 (1975).
6. Badurek, G., Rauch, H. & Tuppinger, D. Neutron interferometric double-resonance experiment. *Phys. Rev. A* **34**, 2600–2608 (1986).
7. Storey, P., Tan, S., Collett, M. & Walls, D. Path detection and the uncertainty principle. *Nature* **367**, 626–628 (1994).
8. Englert, B. G., Scully, M. O. & Walther, H. Complementarity and uncertainty. *Nature* **375**, 367–368 (1995).
9. Storey, E. P., Tan, S. M., Collett, M. J. & Walls, D. F. Complementarity and uncertainty. *Nature* **375**, 368 (1995).
10. Wiseman, H. & Harrison, F. Uncertainty over complementarity? *Nature* **377**, 584 (1995).
11. Wiseman, H. M. et al. Nonlocal momentum transfer in welcher weg measurements. *Phys. Rev. A* **56**, 55–75 (1997).
12. Eichmann, U. et al. Young's interference experiment with light scattered from two atoms. *Phys. Rev. Lett.* **70**, 2359–2362 (1993).
13. Cohen-Tannoudji, C. Effect of non-resonant irradiation on atomic energy levels. *Metrologia* **13**, 161–166 (1977); reprinted in Cohen-Tannoudji, C. *Atoms in Electromagnetic Fields* 343–348 (World Scientific, London, 1994).
14. Kunze, S., Dürr, S. & Rempe, G. Bragg scattering of slow atoms from a standing light wave. *Europhys. Lett.* **34**, 343–348 (1996).
15. Kunze, S. et al. Standing wave diffraction with a beam of slow atoms. *J. Mod. Opt.* **44**, 1863–1881 (1997).
16. Bernhardt, A. F. & Shore, B. W. Coherent atomic deflection by resonant standing waves. *Phys. Rev. A* **23**, 1290–1301 (1981).

Acknowledgements. We thank S. Kunze for discussions. This work was supported by the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to G.R. (e-mail: gerhard.rempe@uni-konstanz.de).

A critical window for cooperation and competition among developing retinotectal synapses

Li I. Zhang*, Huizhong W. Tao*, Christine E. Holt†, William A. Harris† & Mu-ming Poo

Department of Biology, University of California at San Diego, La Jolla, California 92093-0357, USA

* These authors contributed equally to this work.

In the developing frog visual system, topographic refinement of the retinotectal projection depends on electrical activity. *In vivo* whole-cell recording from developing *Xenopus* tectal neurons shows that convergent retinotectal synapses undergo activity-dependent cooperation and competition following correlated pre- and postsynaptic spiking within a narrow time window. Synaptic inputs activated repetitively within 20 ms before spiking of the tectal neuron become potentiated, whereas subthreshold inputs activated within 20 ms after spiking become depressed. Thus both the initial synaptic strength and the temporal order of activation are critical for heterosynaptic interactions among convergent synaptic inputs during activity-dependent refinement of developing neural networks.

Electrical activity in the developing nervous system plays a crucial role in the establishment of early nerve connections^{1,2}. In the mammalian visual system, both the formation of ocular dominance columns in the primary visual cortex^{3–5} and the segregation of retinal ganglion axons into eye-specific layers in the lateral geniculate nucleus⁶ depend on electrical activity in the visual pathways. The pattern of activity in the optic nerves seems to serve an instructive role, as synchronous stimulation of optic nerves abolishes the formation of ocular dominance columns, whereas asynchronous stimulation leads to sharp ocular dominance columns⁷. Artificially synchronized activity in the optic nerve also disrupts the development of orientation tuning in the visual cortex⁸. In the visual system of frog, chick and fish, retinal axons use activity-independent mechanisms initially to establish a topographic map,

but the initial map is coarse and terminals from each retinal axon arborize over a large portion of the tectum. During development, the map becomes refined as retinal axons progressively restrict their arborizations to a smaller fraction of the tectum^{9,10}. Topographic refinement of the retinotectal projection also depends on activity patterns, as this process is impaired when retinal activity is blocked or uniformly synchronized by raising the animals in strobe light^{11–14}. Thus, throughout the visual system, the refinement of connections depends on the pattern of activity, but the underlying physiological mechanisms are largely unknown.

We have examined quantitatively the effects of activity patterns on the strength of developing central synapses in the *Xenopus* retinotectal system. *In vivo* whole-cell recordings were made from neurons in the optic tectum of young *Xenopus* tadpoles to monitor changes in the strength of retinotectal synapses following repetitive electrical stimulation of retinal neurons in the contralateral eye. By

† Present address: Department of Anatomy, University of Cambridge, Cambridge CB2 3DY, UK.

studying an early stage of development, when each retinal ganglion cell axon innervates a large population of tectal cells, we were able to examine the role of activity in both cooperative and competitive interactions among convergent retinotectal connections on a single tectum neuron. We found that repetitive correlated spiking of pre- and postsynaptic neurons can result in persistent synaptic potentiation or depression, depending on the temporal order of the activity among various synaptic inputs. A synaptic input becomes potentiated if it is repetitively activated within 20 ms before spiking of the postsynaptic neuron, and becomes depressed if it is repetitively activated within 20 ms after postsynaptic spiking. These findings indicate the importance of the temporal order of synaptic activation among convergent inputs and set a stringent constraint on the

effective pattern of retinal inputs for the refinement of developing retinotectal connections.

Retinotectal projection in developing brain

Whole-cell perforated patch recordings^{15,16} were made from tectal neurons after surgical exposure of the inner surface of the tectum in *Xenopus* tadpoles (stage 40–41). Extracellular stimulation was applied to the surface of the contralateral retina by two loose-patch electrodes after removal of the lens (Fig. 1a). Evoked excitatory postsynaptic currents (EPSCs) were mediated by the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) subtype of glutamate receptors, as they were abolished by 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX, Fig. 1b). Polysynaptic responses

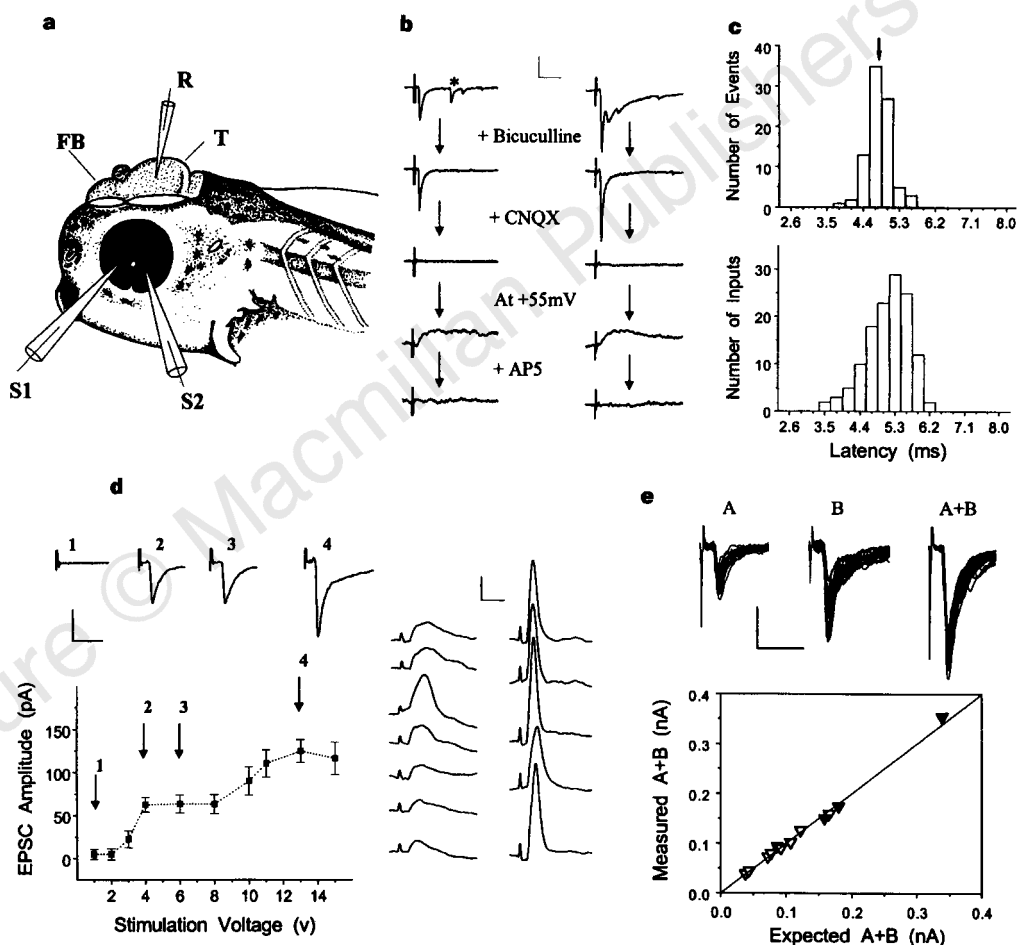


Figure 1 *In vivo* whole-cell recording from *Xenopus* tadpole brain. **a**, Recording arrangement. R, recording electrode; T, tectum; FB, forebrain; S1 and S2; extracellular stimulation electrodes. **b**, Typical evoked synaptic currents recorded in the tectal cell ($V_h = -70$ mV). Left, majority events, a single fast inward current with an onset latency of about 5 ms. Right, minority events, a fast inward current followed by a slow inward current. Each trace is an average of 4 consecutive events; the asterisk marks spontaneous activity. Recordings are from the same cell following sequential addition of bicuculline ($10 \mu\text{M}$), CNQX ($10 \mu\text{M}$) and AP5 ($50 \mu\text{M}$) as indicated, with V_h changed to +55 mV following the CNQX treatment. Scale bars: 30 pA, 25 ms. **c**, Latency of onset of EPSCs. Latency was defined as the time between the onset of 1-ms stimuli and the synaptic current. Top, latency distribution for EPSCs recorded from a tectal cell (arrow marks the mean). Bottom, distribution of the mean latency of a large population of synaptic inputs ($n = 129$). **d**, Dependence of the EPSC amplitude on the stimulus intensity. Data are

amplitudes of EPSCs (mean $\pm$ s.e.m.) induced by stimulation at different voltages, including failures. Arrows 3 and 4 mark the level of 'minimal' and 'non-minimal' stimulation, respectively. Top, average of 10 consecutive events. Scale bars: 50 pA, 15 ms. Right, consecutive traces of the membrane-potential change of the tectal cell following 'minimal' (left) and 'non-minimal' (right) stimulation, respectively. (Scale bars: 15 mV, 15 ms). **e**, Traces are of 15 superimposed, consecutive EPSCs in a tectal neuron evoked by stimulating either one or two converging retinal inputs (A or B) and by co-stimulation of both inputs (A + B), as indicated. Scale bars: 50 pA, 20 ms. The graph shows corresponding values of the calculated sum of mean EPSC amplitudes obtained from separate stimulation (expected A + B) and of measured amplitudes of EPSCs evoked by synchronous co-stimulation of both inputs (measured A + B). Each point represents recording from one tectal cell. Data are from 'minimal' (open symbols) and 'non-minimal' (filled symbols) stimulation, respectively.

observed in some neurons were abolished by bicuculline, indicating the presence of inhibitory postsynaptic currents (IPSCs) mediated by GABA_A (γ -aminobutyric acid A) receptors (Fig. 1b). Consistent with previous reports^{17,18}, these retinotectal synapses contain NMDA (*N*-methyl-D-aspartate) receptors, which were revealed by the appearance of outward EPSCs at V_h of +55 mV after CNQX treatment and the sensitivity of the latter currents to (D)-2-amino-5-phosphovaleric acid (AP5; Fig. 1b). Here we considered EPSCs that had a latency of onset in the range of 3.5 to 6.5 ms following retinal stimulation; these are likely to be monosynaptic in nature. Figure 1c

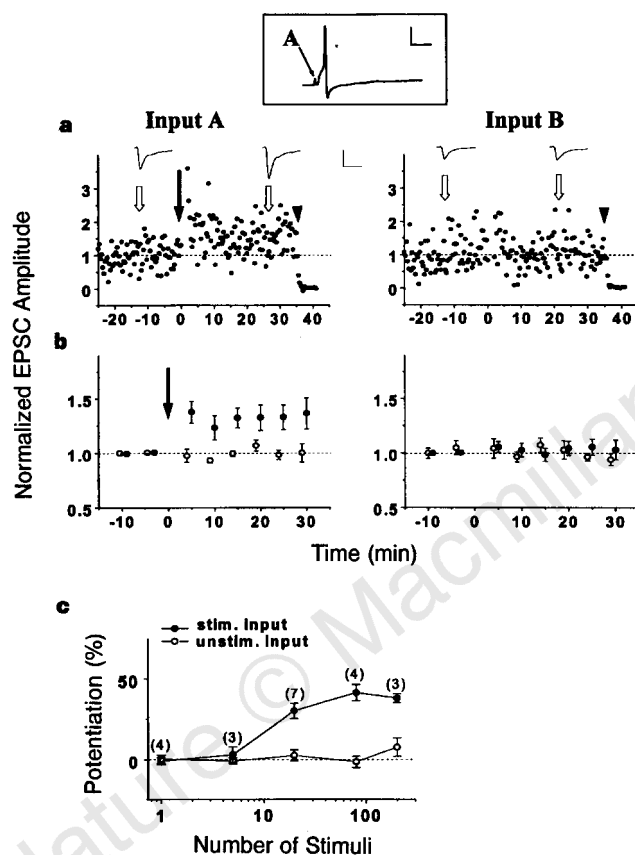


Figure 2 Effects of repetitive stimulation of one of two convergent retinal inputs. **a**, Amplitude of EPSCs elicited by test stimuli (0.063 Hz) at the repetitively stimulated input (A) and unstimulated input (B), respectively, normalized by the mean value observed before repetitive stimulation (dotted line). Sample EPSC traces are the average of 15 consecutive events at the time marked by the arrow. Scale bars: 50 pA, 10 ms. Repetitive stimulation (1 Hz for 20 s at time '0') was applied to input A, resulting in consistent spiking of the tectal cell (in current-clamp, see box; scale bars: 20 mV, 20 ms). Arrowhead indicated CNQX application (10 μ M). **b**, Summary of results from all experiments similar to that shown in **a**. Normalized EPSC amplitudes from each experiment are averaged over 6-min bins before and 5-min bins after repetitive stimulation. Filled circles indicate experiments ($n = 7$) in which the stimulated input A was suprathreshold. The change in the mean EPSC amplitude at the input A 10–30 min after repetitive stimulation (compared with mean pre-stimulation values) was $32 \pm 10\%$ (s.e.m., $n = 7$), whereas that of unstimulated input B was $0.2 \pm 6\%$. Open circles indicate experiments ($n = 4$) in which the stimulated input A was subthreshold (points laterally displaced to the left for clarity). **c**, Dependence of synaptic potentiation on the total number of repetitive stimuli. Potentiation is defined as the percentage increase (mean $\pm$ s.e.m.) in EPSC amplitude 10–30 min after repetitive stimulation. Data are for stimulated (filled circles) and unstimulated (open circles) inputs after different numbers of repetitive stimuli were applied (at 1 Hz). The number of experiments performed is shown in the parenthesis.

shows the distribution of the latency for EPSCs recorded from a typical tectal cell and the distribution of the mean latency for a population of tectal cells studied.

In most experiments, 'minimal' stimulation was used to stimulate the retinal neuron. The stimulus strength was increased in small increments from a low level. EPSCs usually appeared at a threshold stimuli and the mean amplitude remained constant afterwards over a relatively wide range of stimulus intensity (see Fig. 1d). The intensity of the 'minimal' stimulation was set to about 120–150% of the threshold, where a stable 'minimal' EPSC was evoked, consistent with the activation of one or a few neurons. In current clamp, 'minimal' stimulation usually elicited subthreshold evoked potentials (Fig. 1d), although in some cases it was capable of initiating spiking of the tectal neuron. In some experiments, the stimulus intensity was further increased to a level that consistently induced spiking of the tectal cell (Fig. 1d). The latter 'non-minimal' stimulation was likely to activate multiple retinal neurons. Anatomical studies have shown that the projection of single retinal ganglion cell axons in stage 40–41 *Xenopus* tadpoles extends to a large portion of the tectal neuropile^{19,20}. Indeed, EPSCs could be recorded sequentially from several tectal cells along the rostral–caudal axis of the tectum when the same stimulation was applied to the same retinal site (data not shown). When the retina was stimulated at two sites 50–150 μ m apart on the nasal and temporal sides of the optic head (a stimulation configuration we used as standard in this study), there was a relatively high frequency of observing convergent retinal inputs onto a single tectal cell. The evoked responses elicited by two convergent retinal inputs were fully additive, as would be expected if the stimulating electrodes were activating separate populations of retinal ganglion cells (Fig. 1e).

Repetitive stimulation of single inputs

By recording from tectal neurons innervated by two convergent inputs from nasal and temporal parts of the retina, we examined the effect of repetitive stimulation of one of the retinal inputs. The

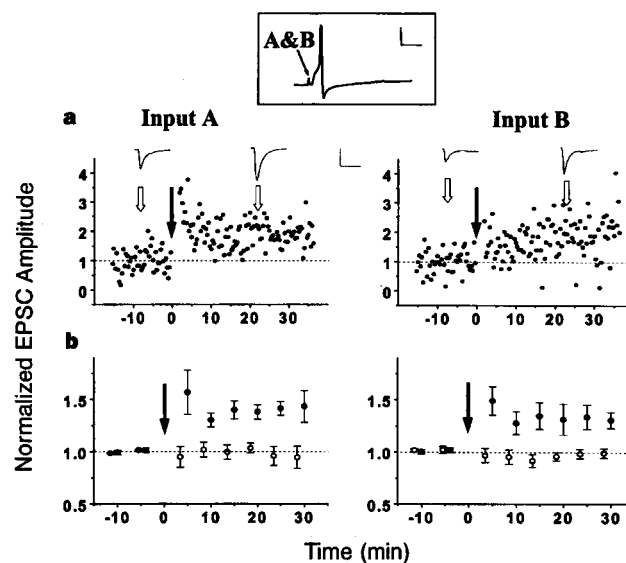


Figure 3 Effects of repetitive synchronous co-stimulation of two convergent retinal inputs. Action potentials were initiated in the tectal cell during repetitive co-stimulation (1 Hz for 20 s). **a**, Results from an experiment in which input A was suprathreshold and input B was subthreshold. **b**, Results from 7 synchronous co-stimulation experiments in which postsynaptic spiking was induced (filled circles). The change in the mean amplitude of EPSCs 10–30 min after repetitive stimulation was $36 \pm 7\%$ (s.e.m., $n = 7$) and $29 \pm 10\%$ (s.e.m.) at inputs A and B, respectively. Open circles show results from 3 experiments in which synchronous co-stimulation did not result in postsynaptic spiking. Scale bars as Fig. 2.

synaptic strength was assayed by measuring the amplitude of EPSCs of each retinal input (A or B) at a low stimulation frequency (0.063 Hz). During the repetitive stimulation of input A (1 Hz for 20 s), the tectal neuron was held in current clamp and action potentials were consistently initiated. As shown in Fig. 2a, the mean EPSC amplitude of input A was increased immediately after the repetitive stimulation of input A, whereas that of the unstimulated input B remained unchanged. Thus repetitive stimulation had resulted in homosynaptic potentiation of the stimulated input without affecting the unstimulated input converging onto the same tectal neuron. A summary of results from seven experiments is shown in Fig. 2b. In a different set of experiments, in which the repetitively stimulated inputs were subthreshold, we found no detectable change in both inputs (Fig. 2b). Thus spiking of the tectal neuron appears to be necessary for the induction of synaptic potentiation by the present protocol. In dually innervated *Xenopus*

myocytes, repetitive stimulation of one input was found to induce persistent heterosynaptic depression of the other unstimulated input²¹, a phenomenon that appears to be absent in this retinotectal system.

To examine further the dependence of synaptic potentiation on the pattern of stimulation, we varied the total number and frequency of repetitive stimuli. The degree of potentiation at the stimulated suprathreshold input increased with an increasing number of stimuli (at 1 Hz) and reached a plateau value at about 80 stimuli (Fig. 2c). No significant change in synaptic strength was observed at unstimulated subthreshold inputs in any of the conditions tested (up to 200 stimuli). Because the unstimulated input was not affected in these experiments following repetitive spiking of the postsynaptic neuron, postsynaptic spiking by itself appears to be insufficient to cause synaptic potentiation. Furthermore, we found that the potentiation effect did not depend on the frequency of

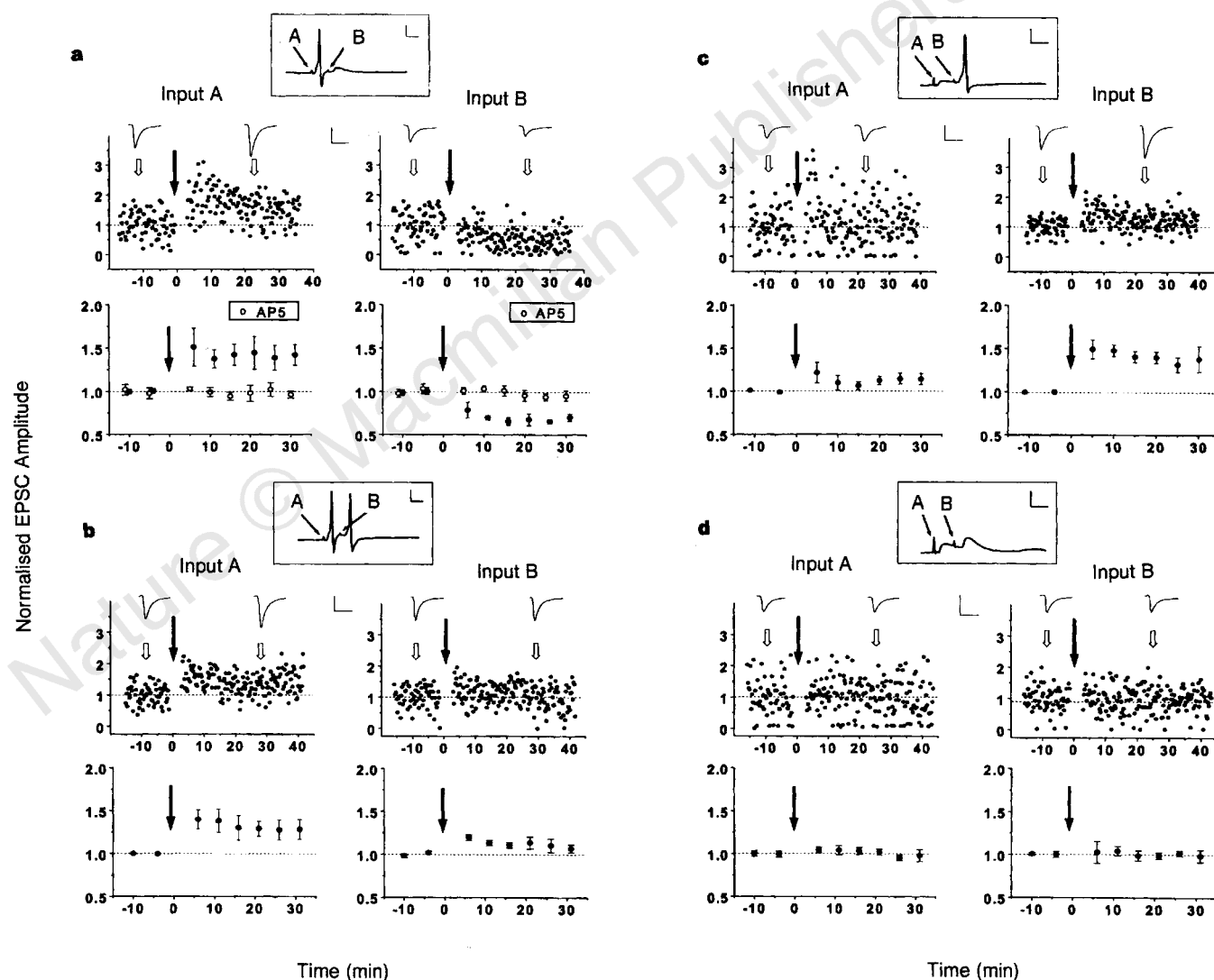


Figure 4 Effects of repetitive asynchronous co-stimulation of two convergent retinal inputs. **a**, Top, examples of synaptic changes induced at two convergent inputs on a tectal cell following repetitive co-stimulation (at time '0', 1 Hz for 100 s), with input A (suprathreshold) stimulated 15 ms before input B (subthreshold), with the tectal cell held in current clamp (see boxed trace; average of 5 events; scale bars: 20 mV, 15 ms). Bottom, results from all 6 experiments similar to that shown above (filled symbols). Note potentiation of input A ($40 \pm 10\%$, $n = 6$) and depression of input B ($-33 \pm 4\%$, $n = 6$). Open symbols show results from 4 experiments identical to that in **a** but with AP5 ($50 \mu\text{M}$) added to the recording

medium. **b**, Same as that in **a** except that both inputs were suprathreshold. Note the marked potentiation of input A ($35 \pm 9\%$, $n = 5$) and much weaker potentiation of input B ($13 \pm 3.6\%$, $n = 5$) following the repetitive stimulation. **c**, As **a**, except that input A was subthreshold and input B was suprathreshold. Note the marked potentiation of input B ($42 \pm 7\%$, $n = 5$) and weaker potentiation of input A ($13 \pm 4.5\%$, $n = 5$). **d**, As **a**, except that both inputs were subthreshold. No change in synaptic strength was observed at either input A ($1.1 \pm 4\%$, $n = 4$) or input B ($0.8 \pm 4\%$, $n = 4$).

stimulation. Application of 100 stimuli at 1, 5 and 20 Hz or of 150 stimuli in theta burst pattern resulted in a similar extent of potentiation of the stimulated input (data not shown). Thus 1 Hz was chosen for our standard stimulation protocol.

Synchronous co-stimulation of convergent inputs

To investigate the issue of synaptic interactions between convergent retinal inputs on a single tectal cell, we examined the effect of repetitive synchronous co-stimulation of both retinal inputs. As shown in Fig. 3a, the amplitude of EPSCs was increased at both inputs following repetitive co-activation for 20 stimuli at a frequency of 1 Hz. In this experiment, input A was suprathreshold, whereas input B was subthreshold. In other experiments, we found similar potentiation of both inputs when both inputs were subthreshold, as long as postsynaptic spiking was consistently initiated. For the seven experiments shown in Fig. 3b, both inputs were subthreshold in two cases (but initiated spiking when co-activated), but in five cases only input B was subthreshold. In two other experiments, in which both inputs were suprathreshold, both inputs became potentiated after repetitive co-stimulation (data not shown). Thus a subthreshold synaptic input became strengthened when it was activated synchronously with a suprathreshold input, but no change was observed if it was not stimulated (Fig. 2b). Consistent with that found for the repetitive stimulation of single inputs, no potentiation of either input was observed when synchronous co-stimulation produced only subthreshold synaptic potential in the tectal cell (Fig. 3b).

Asynchronous co-stimulation of convergent inputs

The effect of temporal pattern of activity on synaptic interactions between converging inputs was explored further by a series of studies using repetitive asynchronous co-stimulation. In the first set of experiments, the first input (A) elicited spiking in the tectal cell, whereas the second input (B), 15 ms later, resulted in only subthreshold EPSPs (Fig. 4a). After 100 paired stimuli at 1 Hz, input A became potentiated and input B became depressed; results from six experiments are summarized in Fig. 4a. In the second set of experiments, both inputs A and B were capable of initiating spiking of the tectal neuron. After the same 100 paired stimuli, input A was markedly potentiated, whereas input B exhibited only a slight potentiation (Fig. 4b). Taken together with the data shown in Fig. 4a, these results suggest that the suprathreshold input B can protect itself from depression induced by the preceding suprathreshold input A. The limited potentiation of input B may be attributed to the depressive effect of the spiking induced by input A. In the third set of experiments, in which the tectal response to input A was subthreshold and input B initiated spiking, input B showed substantial potentiation following repetitive co-stimulation, whereas the potentiation of input A was rather limited (Fig. 4c). In the latter case, the onset of the synaptic response to input A was about 20 ms before the peak of the spike induced by input B, although the interval between the stimuli applied at the retina was 15 ms. Finally, when both inputs were subthreshold, repetitive asynchronous co-stimulation produced no significant effect on the synaptic efficacy of either input (Fig. 4d). These results confirm that postsynaptic spiking is required for the induction of synaptic potentiation. Furthermore, persistent synaptic depression is induced when the subthreshold input is activated within 15–20 ms after spiking of the postsynaptic neuron. Finally, although synaptic potentiation was induced at a subthreshold input when it was activated immediately before postsynaptic spiking (as synchronous co-activation), the potentiation effect largely disappeared when the subthreshold input was activated about 20 ms before the peak of the postsynaptic action potential.

Timing requirements for synaptic modifications

The precise timing of synaptic activation required for the induction

of synaptic potentiation or depression was examined further by varying the time interval between the stimuli repetitively applied to the two converging inputs. We considered the situation illustrated in Fig. 4a, c, in which one retinal input initiated postsynaptic spiking while the other was subthreshold. We varied the interval from –100 to 100 ms and determined the degree of synaptic potentiation or depression for the subthreshold input 10–30 min after the repetitive paired stimulation. Plotting the percentage change in the mean EPSC amplitude with the onset time of the synaptic potential relative to the peak of the action potential elicited in the tectal cell (Fig. 5) allowed us to identify distinct windows for the induction of potentiation and depression. Synaptic potentiation was found for inputs that repetitively arrived within 20 ms before the peak of the action potential, and depression was found for those repetitively arriving within 20 ms after the peak of the action potential. A transition between maximal potential and maximal depression occurs with only a 10-ms change in the timing of the synaptic input.

Cellular mechanisms

Activation of NMDA receptors is required for activity-dependent segregation of retinal axons into eye-specific stripes in frog transplanted with a third eye²² and for the refinement of the topographic map^{23,24}. Activity-induced long-term potentiation (LTP) and depression (LTD) in the CA1 region of the hippocampus^{25,26} and in the visual cortex^{27–29} and lateral geniculate nucleus (LGN)^{30,31} have also been shown to depend on the activation of NMDA receptors. Here we have added AP5 (50 μ M), a selective NMDA receptor antagonist, to the perfusion medium. This treatment did not affect the spiking of the tectal neuron. However, we found that asynchronous paired stimulation similar to that described above for

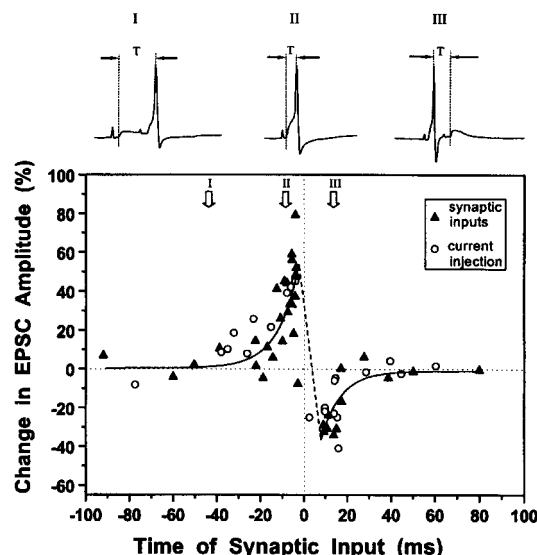


Figure 5 The critical window for synaptic potentiation and depression. The percentage change in the EPSC amplitude of synaptic inputs 10–30 min after repetitive stimulation was plotted against the time of the input (defined by the onset time of the EPSP relative to the peak of the action potential initiated in the tectal cell). Filled triangles show data from experiments similar to those described in Fig. 4a, c. Two converging inputs (one suprathreshold and one subthreshold) were stimulated repetitively (at 1 Hz for 100 s), with varying intervals between the stimuli applied to the two inputs. Only changes in the strength of the subthreshold input were plotted. The results for synaptic inputs were fitted separately for positive and negative times with first-order kinetics, as shown by solid curves. Open circles show data from experiments in which repetitive spiking (at 1 Hz for 100 s) of the tectal cell was induced by injections of depolarizing currents at different times with respect to a subthreshold synaptic input.

Fig. 4a was ineffective in modifying the synaptic strength at either input. The results from five experiments are shown in Fig. 4a. Thus activation of NMDA receptors (see Fig. 1b) is required for the induction of persistent synaptic modification at retinotectal synapses.

Studies of rat brain slices have shown that back-propagated action potentials initiated in the postsynaptic neuron by current injection at the soma exert a critical influence on the strength of synaptic inputs at dendrites^{32–34}. The synaptic input became potentiated when it was activated repetitively 10 ms before the action potential, and was depressed if activated repetitively 10 ms after the action potential³². Our data show that action potentials initiated by synaptic inputs are effective in inducing synaptic modification, and that there is a narrow time window for the induction of potentiation or depression. To further test the sufficiency of postsynaptic spiking in inducing synaptic modifications in this system, we applied brief (1 ms) depolarizing currents into the tectal cell (in current clamp) to initiate repetitive spiking at different times before and after the onset of EPSPs evoked by a subthreshold retinal input. After 100 paired pre- and postsynaptic stimuli at 1 Hz, changes in synaptic efficacy were measured. As shown in Fig. 5, the degree of synaptic potentiation or depression depends critically on the time of onset of EPSPs with respect to the action potentials induced by the current injection, in a manner similar to that found for spiking induced by synaptic inputs. Thus spiking of tectal neuron is a sufficient postsynaptic condition for the induction of synaptic modifications. Furthermore, postsynaptic spiking is necessary for the induction of potentiation at suprathreshold inputs induced by repetitive synaptic activation, as voltage clamping the tectal cell at -70 mV during repetitive stimulation of suprathreshold inputs (1 Hz, 100 s) prevented the induction of potentiation. The percentage change in EPSC amplitude at 10–30 min post-stimulation was $0.1 \pm 4.1\%$ ($\pm$ s.e.m., $n = 4$), but was $41 \pm 11\%$ ($n = 9$) when repetitive stimulation was applied in the current-clamp condition. Finally, we examined the involvement of NMDA receptors in synaptic modifications induced using correlated postsynaptic current injections. For spiking repetitively induced within 20 ms after the onset of synaptic response, the changes in the EPSC amplitude were $-1.5 \pm 4.9\%$ ($n = 4$) in the presence of AP5 ($50 \mu\text{M}$) and $39 \pm 8.0\%$ ($n = 4$) without AP5. For spiking induced within 20 min before the onset of synaptic response, the changes were $4.5 \pm 2.5\%$ ($n = 4$) and $-23 \pm 9.8\%$ ($n = 7$) in the presence and absence of AP5, respectively. These results further support the idea that the same induction mechanism underlies the modification induced by either synaptic inputs or current injections.

The 'depression window' roughly coincides with the after-hyperpolarization following the action potential in the tectal neuron. We tested the possibility that hyperpolarization itself induces depression of a subthreshold input by using hyperpolarization induced by direct injection of negative current. As shown in Fig. 6a, induced hyperpolarization in synchrony with repetitive activation of a subthreshold input had no effect on synaptic strength, but the same input was later depressed significantly when repetitive synaptic activation was preceded by spiking of the tectal neuron induced by repetitive injection of positive currents. In five experiments, the change in the EPSC amplitude averaged $3.1 \pm 5.0\%$ following repetitive synaptic activation in the presence of hyperpolarization, indicating that hyperpolarization by itself is not sufficient for inducing depression in the subthreshold input. The following experiments further confirmed the critical importance of postsynaptic spiking. The spiking of tectal cells induced by suprathreshold inputs was prevented by injections of hyperpolarizing currents during repetitive retinal stimulation. We found that this procedure blocked synaptic potentiation at the suprathreshold input ($-1.3 \pm 2.5\%$, $n = 4$; see Fig. 6b). Furthermore, in the case of asynchronous co-stimulation of two suprathreshold inputs (interval 15 ms), hyperpolarizing currents that repetitively blocked spiking

induced by the second input resulted in depression of the latter ($-25 \pm 4.4\%$, $n = 4$; see Fig. 6c). The same synapse showed potentiation following repetitive stimulation in the absence of induced hyperpolarization. In the case of asynchronous co-stimulation of two subthreshold inputs (interval 15 ms), hyperpolarization of the tectal cell during repetitive stimulation of the second input produced no effect on the synaptic efficacy of either input (data not shown). Thus inhibitory inputs, by hyperpolarizing the tectal cell, may serve to modulate interactions between excitatory synapses.

Discussion

Our results indicate that persistent synaptic modifications in the developing retinotectal system are determined both by the temporal

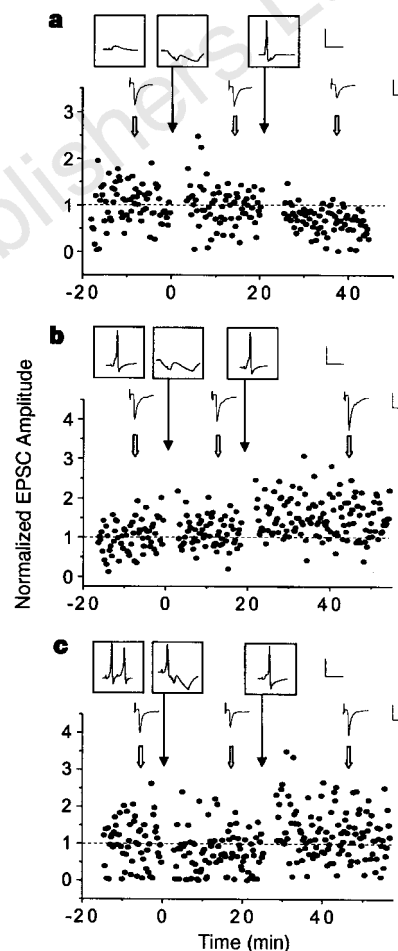


Figure 6 Effects of postsynaptic hyperpolarizations. **a**, Hyperpolarizations (20 mV) were induced in the tectal cell (in current clamp) by injections of negative currents (50 ms duration) in synchrony with repetitive stimulation of a subthreshold input. Boxes above the graphs show sample evoked potentials (average of 5 events; scale bars: 40 mV, 30 ms) during the test period (left) and during repetitive stimulation (black arrows; 1 Hz for 100 s). Lower traces show sample EPSCs (average of 15; scale bars: 30 pA, 20 ms). During the second episode of repetitive stimulation, spiking was initiated repetitively 10 ms before the presynaptic stimulation by injecting depolarizing currents (2 ms). **b**, Hyperpolarizations (25 mV) in synchrony with repetitive stimulation of a suprathreshold input. The postsynaptic spiking was abolished by hyperpolarization (middle box) during the first but not the second episode of repetitive stimulation. **c**, Repetitive hyperpolarizations (40 mV) in synchrony with the second of the two asynchronously activated suprathreshold inputs (with an interval of 15 ms). The blockade of the spiking induced by the second input (middle box) led to depression of the latter. Further repetitive stimulation of the second input alone, in the absence of hyperpolarizations (right box), led to potentiation of the same input.

order of activation of convergent retinal inputs and by their initial synaptic strength. Suprathreshold input that elicits spiking by itself appears to have a competitive advantage among converging inputs, because the onset of its EPSPs is always within the 'potentiation window' preceding the spiking of the postsynaptic neuron. For a subthreshold input, the timing of repetitive activation with respect to other inputs becomes critical in determining its fate: arrival within a 20-ms 'potentiation window' before postsynaptic spiking initiated by other inputs leads to potentiation, whereas arrival within a 20-ms 'depression window' after spiking leads to depression. Suprathreshold inputs can therefore help to strengthen preceding or synchronously activated subthreshold inputs, but can weaken them by initiating spiking immediately before their activation. Subthreshold inputs may cooperate by initiating spiking of the postsynaptic neuron through synchronous co-activation. Among convergent suprathreshold inputs, the input that is repetitively activated earlier than others wins in competition. As shown by the finding that induction of spiking within 20 ms following a preceding spike resulted in a reduced potentiation effect (Fig. 4b), the depression window also seems to apply to suprathreshold inputs. Finally, the effect of hyperpolarization shown in Fig. 6 suggests that convergent inhibitory inputs that are repetitively co-activated synchronously with a suprathreshold input may render the latter a disadvantageous position in synaptic competition, because blocking postsynaptic spiking by hyperpolarization prevents activity-induced potentiation of the suprathreshold input. Thus a retinal input that results in monosynaptic EPSCs followed by one or more polysynaptic GABA-mediated currents (see Fig. 1b), which was observed in 30% of tectal cells recorded, is likely to win in synaptic competition. Local inhibitory circuits established early within the tectum may play an important part in the refinement of retinotectal map.

Increasing the cytosolic Ca^{2+} concentration is known to be critical in the induction of synaptic modification at many synapses^{25,26,35}. It has been suggested that back-propagated action potentials can interact with NMDA receptors to trigger synaptic modifications through Ca^{2+} influx^{32–34}. The critical time windows we have observed may be accounted for by the spatiotemporal pattern of Ca^{2+} elevation in the tectal neuron, resulting from a specific temporal sequence of opening of voltage-dependent Ca^{2+} and NMDA channels. We noted that spiking initiated by a second suprathreshold input does not erase the depressive effect of the first input (see Fig. 4b). Rather, the heterosynaptic depression induced by the first input appears to have reduced the homosynaptic potentiation resulting from the second input.

Of particular relevance to the formulation of Hebb's rule^{36–43}, our results indicate that completely uncorrelated activity and correlated activity of subthreshold inputs that fail to initiate postsynaptic spiking are ineffective in inducing synaptic modifications. Both potentiation and depression depend on correlated pre- and postsynaptic spiking but they require opposite temporal orders in the timing of spikes in the pre- and postsynaptic neurons. This asymmetry in the dependence on the timing of synaptic inputs may be crucial for the synaptic modifications underlying the development and plasticity of neural circuits^{44,45}. A potential mechanism for generating correlated and sequential activation of pre- and postsynaptic neurons is provided by spontaneous waves of activity in the prenatal mammalian retina, which has been shown to be important for segregation of eye-specific layers in LGN^{46–48}. Propagation of waves across the retina leads to the sequential excitation of neighbouring ganglion cells with a defined delay, which may lead to modification of their inputs to a common postsynaptic tectal neuron. In the embryonic or early larval frog, as the visual scene sweeps across the retina it provides correlated and sequential activation of retinal ganglion cells.

During the maturation of the *Xenopus* retinotectal map, a gradual refinement of connections results in the projection of retinal axons

to progressively more restricted populations of tectal cells. In the absence of activity or NMDA-receptor function, this refinement is abolished and axon arbors from retinal ganglion cells increase their spread over the tectum²³. Functional modification of synaptic strength may be a prerequisite or a consequence of the morphological changes associated with the sharpening of the map. The finding that functional modification of synaptic efficacy precedes the removal of nerve terminals during synapse elimination in a developing neuromuscular system⁴⁹ suggests that the structural and functional plasticity of developing nerve connections may be causally related. Because the functional modification of retinotectal synapses and the structural refinement of the retinotectal map depend similarly on the activity and NMDA receptors, the precise activity pattern and timing requirement we observed may help to predict how visual inputs of specific patterns can shape the connectivity in the developing visual system. □

Methods

Xenopus laevis tadpoles were staged using the criteria of Nieuwkoop and Faber⁵⁰. The stage 40–41 tadpole was secured by insect pins to a Sylgard-coated dish and incubated in HEPES-buffered saline containing (in mM): 115 NaCl, 2 KCl, 10 HEPES, 2.5 CaCl_2 , 10 glucose, 1.5 MgCl_2 , 0.005 glycine (pH 7.3). For recording from tectal cells, the skin on top of the head was removed and the brain was split open along the midline to expose the inner surface of the tectum on one side. The lens of the contralateral eye was removed to expose the surface of the retina. A low dose ($1 \mu\text{g ml}^{-1}$) of α -bungarotoxin was applied to the bath to prevent occasional twitching of muscle fibres during the recording. The toxin treatment did not affect retinotectal responses recorded. As shown in Fig. 1b, in the absence of α -bungarotoxin treatment, tectal responses recorded were accounted for entirely by glutamatergic and GABA-ergic transmission. Furthermore, in three experiments performed on tadpoles not treated with the toxin, we observed persistent activity-induced synaptic potentiation as well as synaptic depression, in a manner similar to that described in the text. Perfusion with 50 μM tubocurarine at the end of these experiments resulted in no significant change in both evoked and spontaneous activity.

The method of whole-cell perforated patch recording¹⁵ was used to record from tectal cells, using amphotericin B (Sigma) for perforation¹⁶. The micro-pipettes were made from borosilicate glass capillaries (Kimax), with a resistance in the range 4–6 M Ω . The pipettes were coated with a layer of Sylgard except near the tip, and were tip-filled with internal solution and then back-filled with internal solution containing 200 $\mu\text{g ml}^{-1}$ of amphotericin B. The internal solution contained (in mM): 110 K-gluconate, 10 KCl, 5 NaCl, 1.5 MgCl_2 , 20 HEPES, 0.5 EGTA (pH 7.3). The same patch pipette was used for stimulation of retinal neurons, except that the tip opening was increased to 3 μm , and the filling solution was the bath solution described above. The bath was constantly perfused with fresh recording medium at a slow rate throughout the experiment, and all experiments were performed at room temperature. Recordings were performed with a patch-clamp amplifier (Axopatch 1D; Axon Instruments). The series resistance (15–50 M Ω) was compensated at 75–85% (lag 60 μs). Signals were filtered at 5 kHz using amplifier circuitry and stored on a VCR. Data were sampled at 10 kHz and analysed using pClamp 6.0 software (Axon Instruments). Data accepted for analysis were cases where the average amplitude of EPSCs did not vary beyond 10% of the average value during the control period (15 min), the series resistance did not change by more than 10%, and input resistance (0.5–2 G Ω) remained relatively constant throughout the experiment. During the experiment, the spontaneous activity recorded in tectal cells was monitored continuously by a fast-transient chart recorder (Gould TA240). The stability of amplitude and frequency of spontaneous events was used as an indication of viability of the preparation. For assaying synaptic connectivity, the retinal neuron was extracellularly stimulated by 'blind' loose-patch at a low frequency (0.063 Hz) on the surface layer of the retina. Both EPSCs and IPSCs were inward currents at the clamping potential (–70 mV). The IPSCs have distinctly longer decay times and more negative reversal potentials than EPSCs. As shown in Fig. 1b, pharmacological studies indicated that EPSCs observed at negative membrane potentials are mediated by AMPA receptors, blocked by CNQX (10 μM , RBI). NMDA receptors are activated at positive membrane potentials and are blocked by the antagonist AP5 (50 μM ,

RBI). IPSCs are mediated by GABA_A receptors and blocked by bicuculline methiodide (10 μ M, RBI). Owing to variation in the probability of initiating spiking of the tectal cell, experiments were carried out only on those synaptic inputs that were either strong enough to initiate spiking for most (>70%) stimuli (under either 'minimal' or 'non-minimal' stimulation) or rarely (<10%) initiated spiking; these inputs are defined as 'suprathreshold' and 'subthreshold', respectively.

Received 15 May; accepted 29 June 1998.

1. Goodman, C. & Shatz, C. J. Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* **72**, 77–98 (1993).
2. Katz, L. C. & Shatz, C. J. Synaptic activity and the construction of cortical circuits. *Science* **274**, 1133–1138 (1996).
3. Hubel, D. H. & Wiesel, T. N. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J. Physiol. (Lond.)* **206**, 419–436 (1970).
4. LeVay, S., Hubel, D. H. & Wiesel, T. N. The development of ocular dominance columns in normal and visually deprived monkeys. *J. Comp. Neurol.* **191**, 1–51 (1980).
5. Stryker, M. P. & Harris, W. A. Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *J. Neurosci.* **6**, 2117–2133 (1986).
6. Shatz, C. J. & Stryker, M. P. Prenatal tetrodotoxin infusion blocks segregation of retinogeniculate afferents. *Science* **242**, 87–89 (1988).
7. Stryker, M. P. & Strickland, S. L. Physiological segregation of ocular dominance columns depends on the pattern of afferent electrical activity. *Invest. Ophthalmol. Vis. Sci.* **25**, (Suppl.) 278 (1984).
8. Weliky, M. & Katz, L. C. Disruption of orientation tuning in visual cortex by artificially correlated neuronal activity. *Nature* **386**, 680–685 (1997).
9. Udin, S. B. & Fawcett, J. W. Formation of topographic maps. *Annu. Rev. Neurosci.* **11**, 289–297 (1990).
10. Holt, C. E. & Harris, W. A. Position, guidance and mapping in the developing visual system. *J. Neurobiol.* **24**, 1400–1422 (1993).
11. Harris, W. A. The effects of eliminating impulse activity on the development of the retinotectal projection in salamanders. *J. Comp. Neurol.* **194**, 303–317 (1980).
12. Reh, T. A. & Constantine-Paton, M. Eye-specific segregation requires neural activity in three-eyed *Rana pipiens*. *J. Neurosci.* **5**, 1132–1143 (1985).
13. Schmidt, J. T. & Buzzard, M. Activity-driven sharpening of the retinotectal projection in goldfish: development under stroboscopic illumination prevents sharpening. *J. Neurobiol.* **24**, 384–399 (1993).
14. Brickley, S. G., Dawes, E. A., Keating, M. J. & Grant, S. Synchronizing retinal activity in both eyes disrupts binocular map development in the optic tectum. *J. Neurosci.* **18**, 1491–1504 (1998).
15. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Arch.* **391**, 85–100 (1981).
16. Rae, J., Cooper, K., Gates, P. & Watsky, M. Low access resistance perforated patch recordings using amphotericin B. *J. Neurosci. Methods* **37**, 15–26 (1991).
17. Wu, G.-Y., Malinow, R. & Cline, H. T. Maturation of a central glutamatergic synapse. *Science* **274**, 972–976 (1996).
18. Hickmott, P. W. & Constantine-Paton, M. The contribution of NMDA, non-NMDA, and GABA receptors to postsynaptic responses in neurons of the optic tectum. *J. Neurosci.* **13**, 4339–4353 (1993).
19. Sakaguchi, D. S. & Murphey, R. K. Map formation in the developing *Xenopus* retinotectal system: an examination of ganglion cell terminal arborizations. *J. Neurosci.* **5**, 3228–3245 (1985).
20. O'Rourke, N. A. & Fraser, S. E. Dynamic changes in optic fiber terminal arbors lead to retinotopic map formation: an *in vivo* confocal microscopic study. *Neuron* **5**, 159–171 (1990).
21. Lo, Y. & Poo, M.-m. Activity-dependent synaptic competition *in vitro*: Heterosynaptic suppression of developing synapses. *Science* **254**, 1019–1022 (1991).
22. Cline, H. T., Debski, E. & Constantine-Paton, M. N-methyl-D-aspartate receptor antagonists desegregate eye-specific stripes. *Proc. Natl Acad. Sci. USA* **84**, 4342–4345 (1987).
23. Cline, H. T. & Constantine-Paton, M. NMDA receptor antagonists disrupt the retinotectal topographic map. *Neuron* **3**, 413–426 (1989).

24. Schmidt, J. T. Long-term potentiation and activity-dependent retinotopic sharpening in the regenerating retinotectal projection of goldfish: common sensitive period and sensitivity to NMDA blockers. *J. Neurosci.* **10**, 233–246 (1990).
25. Bliss, T. V. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**, 31–39 (1993).
26. Malenka, R. C. Synaptic plasticity in the hippocampus: LTP and LTD. *Cell* **78**, 535–538 (1994).
27. Artola, A. & Singer, W. Long-term potentiation and NMDA receptors in rat visual cortex. *Nature* **330**, 649–652 (1987).
28. Komatsu, Y., Fujii, K., Maeda, J., Sakaguchi, H. & Toyama, K. Long-term potentiation of synaptic transmission in kitten visual cortex. *J. Neurophysiol.* **59**, 124–141 (1988).
29. Bear, M. F., Press, W. A. & Connors, B. W. Long-term potentiation in slices of kitten visual cortex and the effects of NMDA receptor blockade. *J. Neurophysiol.* **67**, 1–11 (1992).
30. Hahm, J. O., Langdon, R. B. & Sur, M. Disruption of retinogeniculate afferent segregation by antagonists to NMDA receptors. *Nature* **351**, 568–570 (1991).
31. Mooney, R., Madison, D. V. & Shatz, C. J. Enhancement of transmission at the developing retinogeniculate synapse. *Neuron* **10**, 815–825 (1993).
32. Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
33. Magee, J. C. & Johnston, D. A synaptically controlled, associative signal for hebbian plasticity in hippocampal neuron. *Science* **275**, 209–213 (1997).
34. Debanne, D., Gähwiler, B. H. & Thompson, S. M. Long-term synaptic plasticity between pairs of individual CA3 pyramidal cells in rat hippocampal slice cultures. *J. Physiol.* **507**, 237–247 (1998).
35. Naveu, D. & Zucker, R. S. Postsynaptic levels of $[Ca^{2+}]_i$ needed to trigger LTD and LTP. *Neuron* **16**, 619–629 (1996).
36. Hebb, D. H. *The Organization of Behavior* (Wiley, New York, 1949).
37. Stent, G. S. A physiological mechanism for Hebb's postulate of learning. *Proc. Natl Acad. Sci. USA* **70**, 997–1001 (1973).
38. Willshaw, D. J. & Von der Malsburg, C. How patterned neural connections can be set up by self-organization. *Proc. R. Soc. Lond. B* **194**, 431–445 (1976).
39. Bear, M. F., Cooper, L. N. & Ebner, F. F. A physiological basis for a theory of synaptic modification. *Science* **237**, 42–48 (1987).
40. Miller, K. D., Keller, J. B. & Stryker, M. P. Ocular dominance column development: analysis and simulation. *Science* **245**, 605–615 (1989).
41. Constantine-Paton, M., Cline, H. T. & Debski, E. Patterned activity, synaptic convergence, and the NMDA receptor in developing visual pathways. *Annu. Rev. Neurosci.* **13**, 129–154 (1990).
42. Brown, T. H., Kairiss, E. W. & Keenan, C. L. Hebbian synapses: biophysical mechanisms and algorithms. *Annu. Rev. Neurosci.* **13**, 475–511 (1990).
43. Frégnac, Y., Burke, J. P., Smith, D. & Friedlander, M. J. Temporal covariance of pre- and postsynaptic activity regulates functional connectivity in the visual cortex. *J. Neurophysiol.* **71**, 1403–1421 (1994).
44. Mehta, M. R., Barnes, C. A. & McNaughton, B. L. Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl Acad. Sci. USA* **94**, 8918–8921 (1997).
45. Gerstner, W. & Abbott, L. F. Learning navigational maps through potentiation and modulation of hippocampal place cells. *J. Comp. Neurosci.* **4**, 79–94 (1997).
46. Meister, M., Wong, R. O., Baylor, D. A. & Shatz, C. J. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science* **252**, 939–943 (1991).
47. Wong, R. O., Chernjavsky, A., Smith, S. J. & Shatz, C. J. Early functional neural networks in the developing retina. *Nature* **374**, 716–718 (1995).
48. Penn, A. A., Riquelme, P. A., Feller, M. B. & Shatz, C. J. Competition in retinogeniculate patterning driven by spontaneous activity. *Science* **279**, 2108–2112 (1998).
49. Colman, H., Nabekura, J. & Lichtman, J. W. Alterations in synaptic strength preceding axon withdrawal. *Science* **275**, 356–361 (1997).
50. Nieuwkroop, P. D. & Faber, J. *Normal Table of Xenopus laevis* 2nd edn (North Holland, Amsterdam, 1967).

Acknowledgements. We thank G.-q. Bi, B. Berninger, Y. Dan, S. McFarlane and H. Cline for discussion and comments. This work was supported by grants from NIH, NSF and MRC.

Correspondence and requests for materials should be addressed to M.-m.P. (e-mail: mpoo@ucsd.edu).

Constraints on the Hubble constant from observations of the brightest red-giant stars in a Virgo-cluster galaxy

William E. Harris*, Patrick R. Durrell†, Michael J. Pierce‡ & Jeff Secker§

* Department of Physics & Astronomy, McMaster University, Hamilton, Ontario L8S 4M1, Canada

† Department of Astronomy, Case Western Reserve University, Cleveland, Ohio 44106, USA

‡ Department of Astronomy, Indiana University, Bloomington, Indiana 47405, USA

§ Program in Astronomy, Washington State University, Pullman, Washington 99164-3113, USA

The nearest large groups of elliptical galaxies, in the Virgo and Fornax clusters, play a central role in determinations of the Hubble constant, H_0 , and hence the cosmological rate of expansion. Because the relative distances between these two clusters and more remote clusters are well known, absolute distance determinations to Virgo and Fornax should establish the Hubble constant for the local Universe. In addition, elliptical galaxies reside predominantly in the cores of clusters, so distance calibrations for ellipticals should minimize the uncertainties due to the possibly large extent of the clusters along the line of sight. A powerful and direct way of establishing such distances is to use the brightest red-giant stars, which have nearly uniform luminosities^{1,2}. Here we report the direct observation of old red-giant stars in a dwarf elliptical galaxy in the Virgo cluster. We determine a distance to this galaxy, and thus to the core of the Virgo cluster, of 15.7 ± 1.5 megaparsecs, from which we estimate a Hubble constant of $H_0 = 77 \pm 8 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Under the assumption of a low-density Universe with the simplest cosmology, the age of the Universe is no more than 12–13 billion years.

For the old, low-mass stars that dominate the stellar population in elliptical galaxies, the post-main-sequence evolution along the giant branch terminates at the point of core helium ignition which defines the top of the red-giant branch (TRGB). If these stars are imaged in the near-infrared I band, the small dependence of the TRGB bolometric luminosity on stellar composition (metallicity) almost completely cancels the bolometric correction, leaving the absolute magnitude M_I^{TRGB} at the same level to within ± 0.1 magnitude for any metallicity in the range $[\text{Fe}/\text{H}] \approx -0.7$ (refs 1–4). The TRGB luminosity can be calibrated from the giant branches in nearby globular clusters^{1,5}, whose distances are set in turn by RR Lyrae parallaxes and main-sequence fitting to subdwarfs. The TRGB method thus yields distance measurements for other galaxies that are only two steps removed from fundamental trigonometric parallaxes and independent of any other methods involving Local Group calibrators. By contrast, the Cepheid-based distance calibrations^{6–10} lie three or four steps away from fundamental parallax methods by the time they reach remote galaxies such as the rich proving grounds of Virgo and Fornax.

We have used the Hubble Space Telescope (HST) to achieve a TRGB distance calibration to a Virgo galaxy. Our target was the nucleated dwarf elliptical VCC1104 (= IC3388), which has a projected location $43'$ (or $\sim 200 \text{ kpc}$) northeast of the central giant M87, and thus well within the $\sim 1\text{-Mpc}$ projected core radius of the Virgo ellipticals¹¹. We obtained deep I-band (filter F814W) exposures of this galaxy in June 1997 with the WFPC2 camera on the HST. Twelve images adding up to a total exposure time of 32,200 s were re-registered and co-added to generate a final composite image

in which the galaxy is well resolved, displaying a sheet of faint stars across the entire WFPC2 field.

Photometry of the stars on the final image was carried out with the DAOPHOT II suite of software¹². To set the zero point of the I-band magnitude scale, we used our ground-based WIYN (Wisconsin, Indiana, Yale, National Optical Astronomy Observatories) photometry to determine I-band magnitudes for six of the brightest stars in the WFPC2 image field, and then determined the mean offset between the instrumental magnitudes and the true I-band magnitude (I) to within an uncertainty $\sigma(I) = \pm 0.04$. The normal prescription for self-calibration of the WFPC F814W filter via aperture photometry¹³ was also used on 14 stars in the WF2,3,4 fields, yielding a zero point with $\sigma(I) = \pm 0.03$ uncertainty. These two methods gave zero points which differed by only 0.02 mag. An additional systematic zero-point uncertainty may arise when we transfer the magnitude scale for the relatively bright calibration stars to the faint red-giant-branch (RGB) stars we are interested in, because of the charge-transfer efficiency (CTE) effect on the WFPC2 CCDs¹⁴. The CTE effect should, however, be relatively small here because our individual (full-orbit) exposures are the longest possible, yielding background levels of 25 data numbers per exposure. The published CTE equations¹⁴ suggest that any zero-point correction is likely to be no larger than ~ 0.04 mag, but it should be kept in mind when interpreting the total error budget listed in Table 1.

We used additional ground-based images of VCC1104 (obtained with the WIYN telescope at Kitt Peak National Observatory) to measure its integrated colour profiles and thus its metallicity. We find $(U - B) = 0.20 \pm 0.05$ and $(B - V) = 0.74 \pm 0.02$, which give $[\text{Fe}/\text{H}] \approx -1.3 \pm 0.2$ through the normal relation between metallicity and integrated colour for globular clusters¹⁵. The galaxy is therefore in the low-metallicity regime within which M_I^{TRGB} is constant.

Close to the centre of VCC1104, individual stars are too crowded for reliable photometry, even with the high resolution of the HST. We conservatively kept only measurements of stars more than $15''$ (equivalent to $\sim 1 \text{ kiloparsec}$ in projected distance) from the galaxy centre, leaving a total 'clean' sample of $\sim 1,500$ stars brighter than $I = 27.5 \text{ mag}$ for which blending and crowding were negligible. Extensive artificial-star simulations show that our limiting

Table 1 Virgo distance error budget

Type of uncertainty	$\sigma(m - M)$ (mag)
TRGB	
WFPC2 photometric zeropoint uncertainty	± 0.03
TRGB uncertainty from LF model fit	± 0.06
Uncertainty in Virgo foreground extinction	± 0.02
Internal scatter in M_I^{TRGB} calibration	± 0.10
Uncertainty in globular cluster distance scale	± 0.10
Geometric depth of Virgo core ($\pm 1 \text{ Mpc}$)	± 0.13
Total uncertainty for TRGB technique (1 galaxy)	± 0.20
Cepheids	
$P-L$ relation fit of LMC to Virgo spiral	± 0.17
Uncertainty in Cepheid metallicity	± 0.05
Uncertainty in adopted LMC distance	± 0.10
Uncertainty in Virgo extinction relative to LMC	± 0.02
Geometric depth of Virgo spirals ($\pm 3 \text{ Mpc}$)	± 0.35
Total uncertainty for Cepheid technique (1 galaxy)	± 0.41
Planetary nebulae	
Fitting uncertainty to observed PNLF	± 0.10
Photometric zero point and filter calibration uncertainties	± 0.06
Uncertainty in foreground Virgo extinction relative to M31	± 0.02
Uncertainty in fiducial M31 distance	± 0.10
Definition in fiducial model PNLf	± 0.05
Geometric depth of Virgo core ($\pm 1 \text{ Mpc}$)	± 0.13
Total uncertainty for PNLf technique (1 galaxy)	± 0.21

A quantitative overview of the principal internal and external uncertainties in determining the Virgo distance modulus. Figures shown are for one Virgo galaxy picked at random, either from the Virgo E and dE population (for the TRGB and PNLf techniques), or from the spirals (for the Cepheid technique). Data for the PNLf technique are taken from ref. 18, and from ref. 8 for the Cepheids. LF, luminosity function; LMC, Large Magellanic Cloud; $P-L$, period-luminosity.

magnitude (50% detection completeness) is $I \approx 27.5$ mag in these outer regions.

The raw luminosity function of the detected stars is shown in the upper panel of Fig. 1. It has three components: (1) the dominant contribution of the RGB stars in the galaxy; (2) a small ($\leq 20\%$) contribution of brighter asymptotic giant branch (AGB) stars on their second giant branch ascent; and (3) a contribution from faint, star-like background objects. We model the entire luminosity function through a maximum-likelihood fitting approach¹⁶. The RGB luminosity function $N(I)$ is assumed to have a step-function onset at $I(\text{TRGB})$, followed by an exponential rise to fainter magnitudes. To this we add an AGB population starting 1.0 mag brighter than the TRGB and rising exponentially to fainter magnitudes (though with much smaller amplitude than the RGB itself). We set the exponential slopes for both RGB and AGB at $\Delta \log N / \Delta I = 0.96$ from the $(\log N, I)$ form of the luminosity function. Finally, we convolve the ideal model with the photometric error function $\sigma(I)$ and completeness function $f(I)$ (both of which are known from the artificial-star simulations). The smoothed

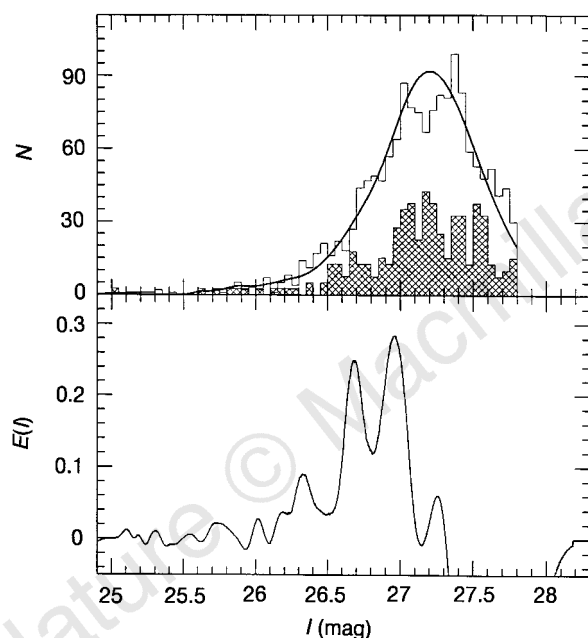


Figure 1 Luminosity function. Top, the observed luminosity function $N(I)$ (number of stars per 0.1-mag bin) for the red giants in VCC1104. The solid line superimposed on the raw data shows the scaled maximum-likelihood fit of the model luminosity function (LF) to the data, scaled to match the total population of stars and including the effects of background contamination, photometric error, and detection completeness. The steep rise starting near $I \approx 26.7$ mag shows the tip of the RGB; ideally, it would show up as a step-function jump exactly at the RGB tip, but it is broadened out by the photometric measurement uncertainty, which is $\sigma(I) = \pm 0.15$ mag at the TRGB level and increases steadily to fainter levels. The steep fall-off past $I \approx 27.4$ mag is due to the declining completeness of detection of faint stars, which is nearly 100% at the TRGB but drops to 50% at $I \approx 27.5$ mag and effectively to zero at $I = 28$ mag. The small bump in the LF at the bright end ($I \approx 26.3$ mag) is likely to be due to the presence of bright AGB stars. The local background function b —the raw, unsmoothed form of which is shown as the hatched histogram—is defined empirically from the LF on the outermost half of the WF4 camera field, which is furthest from the centre of VCC1104 and in which the stellar density is lowest. However, even at the outer edges of the field, the RGB population from the galaxy contributes noticeably to b and thus the ‘true’ background level is significantly lower than is shown here. Bottom, results of the Sobel edge-detection algorithm, as applied to the smoothed and normalized LF. The function $E(I)$ shows a sharp peak wherever it passes through a rapid change in the LF. The AGB population shows up as a small ‘edge’ at $I \approx 26.3$ mag, but it is the peak at $I \approx 26.7$ mag which we identify as the RGB tip.

background $b(I)$ is added, and the fully transformed model is finally matched to the observed luminosity function. In general, the luminosity function for this galaxy very strongly resembles those of two dwarf ellipticals in the nearby M81 group¹⁷, for which the $(I, V - I)$ colour–magnitude diagrams exhibit a strong RGB component terminating sharply at the TRGB, with a sprinkling of AGB stars continuing upward brighter than the tip for ~ 0.8 mag.

The free parameters in the luminosity function fit are the assumed I_{TRGB} and the fractional contribution of the AGB, which are adjusted to find the best match (maximum likelihood) to the normalized luminosity function. Our model solution gives $I_{\text{TRGB}} = 26.82 \pm 0.06$. Although variations in the size of the AGB population, or differences in the RGB slope exponent do affect the overall goodness of fit of the model luminosity function, the deduced magnitude of the TRGB is almost completely insensitive to either parameter because of its intrinsically sharp step-function nature. (Realistic changes in the exponential slope, or the AGB population ratio, resulted in only small ($\pm 0.01 - 0.02$) differences in the TRGB value). Similarly, even neglecting the contribution of the background $b(I)$ altogether does not change the deduced TRGB significantly. As a check on the solution, we applied the Sobel edge-detection filter¹³ in its continuous form to the observed luminosity function (lower panel of Fig. 1). This method, though cruder, has the advantage that it is model-independent; it gives the same answer for I_{TRGB} to within ± 0.1 mag.

The distance to the galaxy is set once we adopt an absolute magnitude for the TRGB, $M_I^{\text{TRGB}} = -4.2 \pm 0.1$ (Fig. 2). Our derived distance modulus for VCC1104 is then $(m - M)_0 = 30.98 \pm 0.20$ for a foreground Virgo absorption $A_I = 0.04$. Our resulting distance to VCC1104, and therefore the Virgo cluster core, is $d_{\text{Virgo}} = 15.7 \pm 1.5$ Mpc. This result agrees to within its quoted uncertainty with other recent measurements such as those determined by planetary nebula luminosity functions¹⁸, surface brightness fluctuations for E/S0 galaxies¹⁹, and the Cepheids in spiral galaxies that are high-probability Virgo members^{6–10}.

The error budget is summarized in Table 1 along with those for the Cepheid and planetary nebula luminosity function (PNLF) techniques. For single Virgo galaxies selected at random, a significant advantage of either the TRGB or PNLf methods should be a smaller geometric depth effect, because the Virgo ‘core’ defined by the elliptical galaxies is expected to be smaller than the more widespread Cepheid spirals (but see below).

Two extreme possibilities may have significantly biased our result. (1) We may have unluckily picked a target which genuinely lies on the near (or far) side of the Virgo cluster core. The distribution of

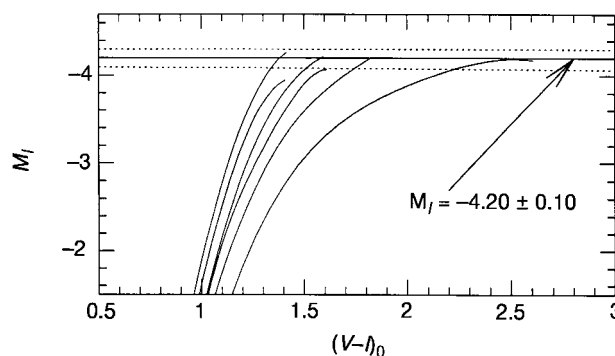


Figure 2 Fiducial giant branches for standard Milky Way globular clusters (absolute magnitude M_I against colour $(V - I)_0$). The clusters measured by Da Costa and Armandroff⁵ have been recalibrated through the Hipparcos subdwarf parallax distance scale³⁰. The sequence of cluster RGBs shown covers the metallicity range $[\text{Fe}/\text{H}] = -2.2$ for M15 (leftmost line) up to $[\text{Fe}/\text{H}] = -0.7$ for 47 Tucanae (rightmost line); over this range, the RGB tip is virtually constant at $M_I^{\text{TRGB}} = -4.2 \pm 0.1$. Encouragingly, recent theoretical RGB models (ref. 31) agree well with the empirical calibration.

intergalactic planetary nebula within Virgo²⁰ suggests that the intracluster stars may extend as much as 3 Mpc towards us along the line of sight, though it is unclear whether or not the core distribution E and dE galaxies is similarly extended¹¹. Ultimately, the only safeguard we have against this type of accident is to acquire data for more galaxies of similar type and determine their distribution of distances. (2) This particular dwarf elliptical could have a huge population of bright AGB stars which mimic the normal first-ascent RGB tip, and which would cause us to underestimate its distance. However, no such AGB population is seen in the outskirts of similar dwarf ellipticals (see ref. 17 for comparison), or even in the haloes of larger galaxies^{21–23}.

To compute the Hubble constant, we use our measured Virgo distance to step outwards to the Coma cluster, whose mean redshift in the Cosmic Microwave Background frame^{24,25} is $v_{\text{Coma}} = (7,100 \pm 100) \text{ km s}^{-1}$. The relative distance modulus $\Delta m(\text{Coma} - \text{Virgo})$ appears now to be the most poorly known quantity in the logical chain, with measured estimates ranging from ~ 3.4 to 4.1 (ref. 26). Recent measurements, particularly through the $D_n - \sigma$ method for fundamental-plane ellipticals^{27,28}, suggest a mean $(\Delta m) \approx 3.85 \pm 0.15$, which we adopt here. Thus, $d_{\text{Coma}} = (92.5 \pm 10.3) \text{ Mpc}$ and the Hubble constant is $H_0 = (77 \pm 9) \text{ km s}^{-1} \text{ Mpc}^{-1}$. The distance to Coma that we find is consistent with the Leo–Coma relative distance derived from the TRGB technique on NGC 3379²³. The corresponding expansion age for the Universe is $(12.6 \pm 1.5) \text{ Gyr}$ for a cosmological density parameter Ω of ~ 0 , or $8.4 \pm 1.0 \text{ Gyr}$ for critical density ($\Omega = 1$). The former value, for a very low-density universe, now fits, with little room to spare, within the lowest currently estimated ages for the oldest Milky Way globular clusters, which are in the range 11–13 Gyr (refs 29, 30). The expansion time for the high-density universe is, of course, still strongly in conflict with these ages. □

Received 17 February; accepted 9 June 1998.

1. Lee, M. G., Freedman, W. L. & Madore, B. F. The tip of the red giant branch as a distance indicator for resolved galaxies. *Astrophys. J.* **417**, 553–559 (1993).
2. Mould, J. & Kristian, J. The stellar populations in the halos of M31 and M33. *Astrophys. J.* **305**, 591–599 (1986).
3. Lee, M. G. The distance to nearby galaxy NGC 3109 based on the tip of the red giant branch. *Astrophys. J.* **408**, 409–415 (1993).
4. Sakai, S., Madore, B. F. & Freedman, W. L. Tip of the red giant branch distances to galaxies. III. The dwarf galaxy sextans A. *Astrophys. J.* **461**, 713–723 (1996).
5. Da Costa, G. S. & Armandroff, T. E. Standard globular cluster giant branches in the $(M_I, (V - I)_0)$ plane. *Astron. J.* **100**, 162–181 (1990).
6. Pierce, M. J. *et al.* The Hubble constant and Virgo cluster distance from observations of Cepheid variables. *Nature* **371**, 385–389 (1994).
7. Freedman, W. L. *et al.* Distance to the Virgo cluster galaxy M100 from Hubble Space Telescope observations of Cepheids. *Nature* **371**, 757–762 (1994).
8. Ferrarese, L. *et al.* The extragalactic distance scale key project. IV. The discovery of Cepheids and a new distance to M100 using the HST. *Astrophys. J.* **464**, 568–599 (1996).
9. Saha, A. *et al.* Cepheid calibration of the peak brightness of SNeIa. V. SN1981B in NGC 4536. *Astrophys. J.* **466**, 55–91 (1996).
10. Saha, A. *et al.* Cepheid calibration of the peak brightness of SNeIa. VI. SN1960F in NGC 4496A. *Astrophys. J. Suppl.* **107**, 693–737 (1996).
11. Ferguson, H. C. & Sandage, A. The spatial distributions and intrinsic shapes of dwarf elliptical galaxies in the Virgo and Fornax clusters. *Astrophys. J.* **346**, L53–L56 (1989).
12. Stetson, B., Davis, L. E. & Crabtree, D. R. in *CCDs in Astronomy* (ed. Jacoby, G. H.) 289–304 (ASP Conf. Ser. 8, Astron. Soc. Pacific, San Francisco, 1990).
13. Holtzman, J. A. *et al.* The photometric performance and calibration of WFPC2. *Publ. Astron. Soc. Pac.* **107**, 1065–1093 (1995).
14. Whitmore, B. & Heyer, I. *New Results on Charge Transfer Efficiency and Constraints on Flat-Field Accuracy* (Instrument Science Rep. 97-08, Space Telescope Science Institute, Baltimore, 1997).
15. Couture, J., Harris, W. E. & Allwright, J. W. B. BV1 photometry of globular clusters in M87. *Astrophys. J. Suppl.* **73**, 671–683 (1990).
16. Secker, J. & Harris, W. E. A maximum-likelihood analysis of globular cluster luminosity distributions in the Virgo cluster. *Astron. J.* **105**, 1358–1368 (1993).
17. Caldwell, N., Armandroff, T. E., Da Costa, G. S. & Seitzer, P. Dwarf elliptical galaxies in the M81 group: the structure and stellar populations of BK5N and F8D1. *Astron. J.* **115**, 535–558 (1998).
18. Jacoby, G. H., Ciardullo, R. & Ford, H. C. Planetary nebulae as standard candles. V. The distance to the Virgo cluster. *Astrophys. J.* **356**, 332–349 (1990).
19. Tonry, J. L., Blakeslee, J. P., Ajhar, E. A. & Dressler, A. The SBF survey of galaxy distances. I. Sample selection, photometric calibration, and the Hubble constant. *Astrophys. J.* **475**, 399–413 (1997).
20. Feldmeier, J. J., Ciardullo, R. & Jacoby, G. H. Intracluster planetary nebulae in the Virgo cluster I. Initial results. *Astrophys. J.* (in the press).
21. Soria, R. *et al.* Detection of the tip of the red giant branch in NGC 5128. *Astrophys. J.* **465**, 79–90 (1996).
22. Elson, R. A. W. Red giants in the halo of the S0 galaxy NGC 3115: a distance and a bimodal metallicity distribution. *Mon. Not. R. Astron. Soc.* **286**, 771–776 (1997).
23. Sakai, S. *et al.* Detection of the tip of the red giant branch in NGC 3379 (M105) in the Leo I group using the Hubble Space Telescope. *Astrophys. J.* **478**, 49–57 (1997).

24. Colless, M. & Dunn, A. M. Structure and dynamics of the Coma cluster. *Astrophys. J.* **458**, 435–454 (1996).
25. Mould, J. *et al.* Limits on the Hubble constant from the HST distance of M100. *Astrophys. J.* **449**, 413–421 (1995).
26. van den Bergh, S. The Hubble parameter. *Publ. Astron. Soc. Pacif.* **104**, 861–883 (1992).
27. Faber, S. M. *et al.* Spectroscopy and photometry of elliptical galaxies. VI. Sample selection and data summary. *Astrophys. J. Suppl.* **69**, 763–808 (1989).
28. Hjorth, J. & Tanvir, N. R. Calibration of the fundamental plane zero point in the Leo I group and an estimate of the Hubble constant. *Astrophys. J.* **482**, 68–74 (1997).
29. Harris, W. E. *et al.* NGC 2419, M92, and the age gradient in the galactic halo. *Astron. J.* **114**, 1030–1042 (1997).
30. Gratton, R. G. *et al.* Ages of globular clusters from hipparcos parallaxes of local subdwarfs. *Astrophys. J.* **491**, 749–771 (1997).
31. Cassisi, S. & Salaris, M. The tip of the red giant branch as a distance indicator: theoretical calibration and the value of H_0 . *Mem. Soc. Astron. Ital.* (in the press).

Correspondence and requests for materials should be addressed to W.E.H. (e-mail: harris@physics.mcmaster.ca).

A Cepheid distance to the Fornax cluster and the local expansion rate of the Universe

Barry F. Madore^{*}, Wendy L. Freedman[†], N. Silberman[‡], Paul Hardings[§], John Huchra^{||}, Jeremy R. Mould[¶], John A. Graham[#], Laura Ferrarese[☆], Brad K. Gibson^{||}, Mingsheng Han^{**}, John G. Hoessel^{**}, Shaun M. Hughes^{††}, Garth D. Illingworth^{‡‡}, Randy Phelps[†], Shoko Sakai[‡] & Peter Stetson^{§§}

^{*} NASA/IPAC Extragalactic Database, Infrared Processing and Analysis Center, Jet Propulsion Laboratory, California Institute of Technology, MS 100-22, Pasadena, California 91125, USA

[†] Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, California 91101, USA

[‡] Jet Propulsion Laboratory, California Institute of Technology, MS 100-22, Pasadena, California 91125, USA

[§] Steward Observatory, University of Arizona, Tucson, Arizona 85721, USA

^{||} Harvard-Smithsonian Institute, Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

[¶] Mt Stromlo and Siding Spring Observatories, Institute of Advanced Studies, Private Bag, Weston Creek Post Office, ACT 2611, Australia

[#] Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015, USA

[☆] California Institute of Technology, MS 105-24 Robinson Laboratory, Pasadena, California 91125, USA

^{**} Department of Astronomy, University of Wisconsin, 475 N. Charter Street, Madison, Wisconsin 53706, USA

^{††} Royal Greenwich Observatory, Madingley Road, Cambridge CB3 0HA, UK

^{‡‡} Lick Observatory, University of California, Santa Cruz, California 95064, USA

^{§§} Dominion Astrophysical Observatory, 5071 W. Saanich Road, Victoria, BC, Canada V8X 4M6

Both galaxy distances and velocities are required for the determination of the expansion rate of the Universe, as described by the Hubble constant H_0 . The radial velocities of galaxies arise not just from this expansion but also from random components and large-scale flows. To reach out to distances dominated by the overall cosmic expansion, it is necessary to probe large physical scales where galaxy–galaxy and galaxy–cluster interactions become less important. But accurate distances of nearby galaxies and clusters (commonly measured¹ using Cepheid variable stars) are nevertheless required to calibrate the indirect distance indicators generally used to measure these large scales. Here we report a Cepheid distance of 18.6 ± 1.9 (statistical error) $\pm 1.9 \text{ Mpc}$ (systematic error) for the galaxy NGC1365 in Fornax, a cluster of galaxies in the Southern Hemisphere. We find a value of $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ from Fornax alone, and $73 \text{ km s}^{-1} \text{ Mpc}^{-1}$ from the intervening galaxy flow, each corrected for infall into the Virgo cluster. These values are consistent with the Hubble constant measured in the far field using secondary methods². Our

data support previous suggestions^{3–5} that the local small-scale velocity field has very small scatter ($\sim \pm 70 \text{ km s}^{-1}$).

Both the Virgo and Fornax clusters of galaxies are close enough to be studied in considerable detail; on the other hand, they are sufficiently distant that most of their measured velocity is probably dominated by the expansion of the Universe. The Fornax cluster of galaxies is comparable in distance to the Virgo cluster⁶, but it is found almost opposite to Virgo in the skies of the Southern Hemisphere. Fornax is less rich in galaxies than the Virgo cluster⁷, but it is also substantially more compact (see Fig. 1). As a result of its lower mass, the influence of Fornax on the local velocity field is expectedly less dramatic than that of Virgo; and, because of its compact nature, questions concerning the membership of individual galaxies in Fornax are less problematic, and the back-to-front geometry is far less controversial than in the case of the Virgo cluster complex¹. Clearly, Fornax is an interesting site for a direct test of the local expansion rate if its distance were known. Measurement of distances to nearby galaxies has been considerably enhanced by the availability of the Hubble Space Telescope^{1,8,9} (HST). With the HST, Cepheid distances out to and including the Fornax cluster offer a direct means of estimating the expansion rate of the nearby Universe.

The goals of the HST Key Project on the Extragalactic Distance Scale^{8,9} are much broader than measuring the distances to two nearby clusters; but there are several important reasons to secure a distance to the Fornax cluster. Fornax serves both as a probe of the local expansion velocity field and as a well defined location for calibrating several secondary distance indicators which can be used to reach the essentially unperturbed Hubble flow. To obtain the distance to Fornax, the Key Project is searching for Cepheids in three member galaxies. The first of these, discussed here, is the luminous, barred spiral galaxy, NGC1365.

At least three lines of evidence independently suggest that NGC1365 is a physical member of the Fornax cluster. First, NGC1365 is almost directly along our line of sight to Fornax as a whole, being projected only ~ 70 arcmin from the geometric centre of the cluster, whereas the radius of the cluster is ~ 200 arcmin (ref. 10; see Fig. 1). Second, NGC1365 is also coincident with the Fornax cluster in velocity space. The systemic (heliocentric) velocity V and velocity dispersion σ of the main population of galaxies in the inner six degrees of the Fornax cluster are well defined²: 39 spiral/irregular galaxies with published radial velocities give $V = 1,399 \text{ km s}^{-1}$ and

$\sigma = \pm 334 \text{ km s}^{-1}$, 78 E/S0 galaxies give $V = 1,463 \text{ km s}^{-1}$ with $\sigma = \pm 347 \text{ km s}^{-1}$; and the combined (unweighted) sample of 117 galaxies gives $V = 1,441 \text{ km s}^{-1}$ with $\sigma = \pm 342 \text{ km s}^{-1}$ (see, for comparison, refs 11, 12). We note in passing that the mean velocity of the spirals agrees with the mean for the ellipticals to within 64 km s^{-1} . And specifically, the velocity difference between NGC1365 and the cluster mean is only about two-thirds of the cluster's intrinsic velocity dispersion. Third, NGC1365 sits only 0.02 mag from the central ridge line of the apparent Tully–Fisher relation relative to other cluster members defined by recent studies of Fornax^{11,13}. Taken together, these observations strongly suggest that a distance to NGC136 alone should be representative of the cluster as a whole.

Using the Wide Field and Planetary Camera 2 on HST, we have obtained a set of 12-epoch observations of NGC1365 and have discovered 37 Cepheids using two independent photometry routines, and selected by light-curve quality (see ref. 14 for details). The resulting intensity-weighted visual $\langle V \rangle$ and near-infrared $\langle I \rangle$ period–luminosity relations (Fig. 2) were fitted by χ^2 minimization to the fiducial period–luminosity relation for Large Magellanic

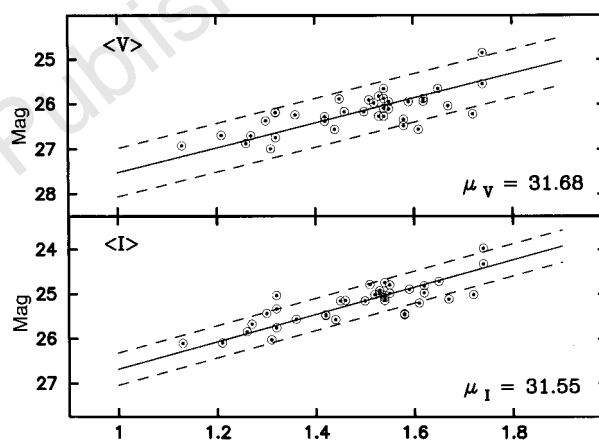


Figure 2 V- and I-band period–luminosity relations for the full set of 37 high-quality Cepheids monitored in NGC1365. The fits are to the fiducial relations¹⁵ shifted to the apparent distance modulus of NGC1365. Dashed lines indicate the expected intrinsic (2σ) width of the relationship due to the finite temperature width of the Cepheid instability strip. The x axis is $\log(\text{period in days})$

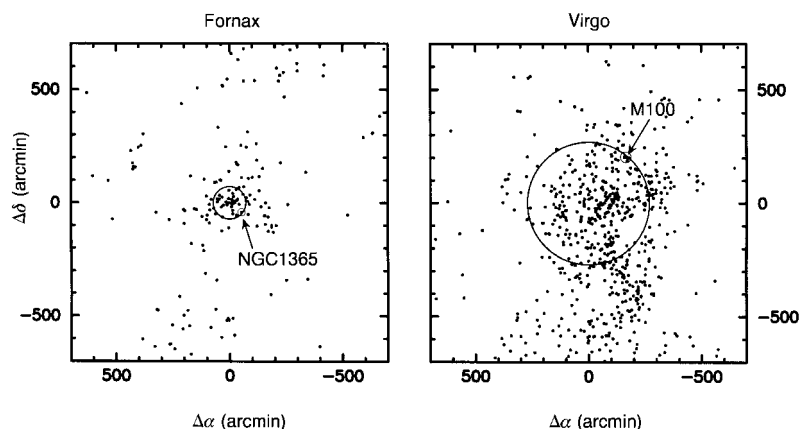


Figure 1 A comparison of the distribution of galaxies, drawn to scale, as projected on the sky for the Virgo cluster (right panel) and the Fornax cluster (left panel). Units are arcmin. The comparison of apparent sizes is appropriate given that the two clusters are at approximately the same distance from us. In the extensive Virgo cluster, the galaxy M100 is marked $\sim 4^\circ$ to the northwest of the elliptical-galaxy-rich core; this corresponds to an impact parameter of 1.3 Mpc, or 8% of the distance from the Local Group to the Virgo cluster. The Fornax cluster is more centrally concentrated than Virgo, so that the back-to-front uncertainty associated

with its three-dimensional spatial extent is reduced for any randomly selected member. Roughly speaking, converting the total angular extent of the Fornax cluster on the sky (6° diameter³) into a back-to-front extent, the error associated with any randomly chosen galaxy in the Fornax cluster translates into a few per cent uncertainty in distance. We anticipate that this uncertainty in distance will be reduced when HST data for two additional Fornax spirals (NGC1425 and NGC1326A) are analysed in the next 12 months.

Cloud (LMC) Cepheids¹⁵ (corrected for a reddening of $E(B - V)_{\text{LMC}} = 0.10$ mag), scaled to an LMC true distance modulus of $\mu_0 = 18.50$ mag and shifted into registration with the Fornax data. The effect of Hipparcos parallaxes on the Cepheid distance scale are still being debated^{16–18}, although several of the more recent discussions conclude that the distance scale through the LMC adopted here is consistent with the new parallax data, with small upward corrections in the distance—and consequently downward corrections in the Hubble constant—being at the 3–5% level. The resulting apparent moduli are $\mu_V = 31.68 \pm 0.05$ mag and $\mu_I = 31.55 \pm 0.05$ mag. Correcting for a total line-of-sight reddening of $E(V - I)_{\text{N1365}} = 0.14$ mag (derived from the Cepheid data themselves, and using a standard reddening law) gives a true distance modulus of $\mu_0 = 31.35 \pm 0.07$ mag for NGC1365. This corresponds to a distance of 18.6 ± 0.6 Mpc. This error quantifies the statistical uncertainty derived from the measured photometric errors, combined with the intrinsic magnitude and colour width of the Cepheid instability strip; a more complete discussion of the errors can be found in ref. 2, giving a cumulative random error of -1.9 Mpc and a systematic uncertainty (quantifying metallicity effects, period–luminosity zero point and photometric uncertainties) also amounting to ± 1.9 Mpc.

Here we point out that, unlike our earlier discussion¹ of the expansion rate derived from a Cepheid distance to the galaxy M100 in the Virgo cluster, the infall-velocity correction for the Local Group motion with respect to the Virgo cluster (and its associated uncertainty) becomes a minor issue for the Fornax cluster. This is the result of a fortuitous combination of geometry and kinematics. The physical separation of the two clusters is readily derived as both their angular separation and their individual distances are now known. Under the assumption that the Virgo cluster dominates the local velocity perturbation field at the Local Group and at Fornax, we can calculate the infall velocity at Fornax (assuming that the flow field amplitude scales with $1/R_{\text{Virgo}}$, characterized by a R^{-2} density distribution¹⁹, where R_{Virgo} is the distance from the centre of the Virgo cluster). From this we then derive the flow contribution to the measured line-of-sight radial velocity, as seen from the Local Group. Figure 3 shows the distance scale geometry (left panel) and the velocity-field kinematics (right panel) of the Local Group–Virgo–

Fornax system. Using an infall velocity⁵ of the Local Group towards Virgo of $+200 \text{ km s}^{-1}$, and adopting a generous uncertainty of $\pm 100 \text{ km s}^{-1}$, the flow correction for Fornax is only $-45 \pm 23 \text{ km s}^{-1}$. This correction and its uncertainty are a factor of more than four times smaller than the equivalent corrections applied to M100.

We are now in a position to derive two estimates of the general expansion rate of the local Universe. The first estimate is based solely on the systematic radial velocity and the Cepheid-based distance to the Fornax cluster, and therefore samples the velocity field in one particular direction at ~ 20 Mpc. We then examine a more isotropic (inner) volume of space, sampled at nine randomly distributed distances and directions out to, and including, both the Virgo and Fornax clusters. This sample has the advantage of providing an average over the volume, but it is still limited in scale (probing an average distance of ~ 10 Mpc). It is also still subject to the uncertainties in the bulk flow of this entire volume, as well as uncertainties in the adopted ‘Virgocentric’ flow model.

The observed velocity of the Fornax cluster, $1,441 \text{ km s}^{-1}$, can be corrected to the barycentre of the Local Group (-90 km s^{-1}) and compensating for the -45 km s^{-1} component of the Virgocentric flow, derived above, to give a cosmological expansion rate of Fornax of $+1,306 \text{ km s}^{-1}$. Using the Cepheid distance of 18.6 Mpc for Fornax then gives $H_0 = 70(\pm 18), [\pm 7], \text{ km s}^{-1} \text{ Mpc}^{-1}$. The first uncertainty (in parentheses) includes global flow uncertainties, random errors in the distance derived from the period–luminosity fit to the Cepheid data, as well as random velocity errors in the adopted Virgocentric flow, combined with the distance uncertainties to Virgo propagated through the flow model. The second uncertainty (in square brackets) quantifies the currently identifiable systematic errors associated with the adopted mean velocity of Fornax, and the adopted zero point of the period–luminosity relation (combining in quadrature the LMC distance error, a measure of the metallicity uncertainty, and a conservative estimate of the photometric errors. An extensive discussion of these errors is given in the discussion and tabulation in ref. 2). Finally, we note that according to the Han–Mould model¹², the so-called Local Anomaly gives the Local Group an extra velocity component of approximately $+73 \text{ km s}^{-1}$ towards Fornax. If we were to add that correction our local estimate, H_0

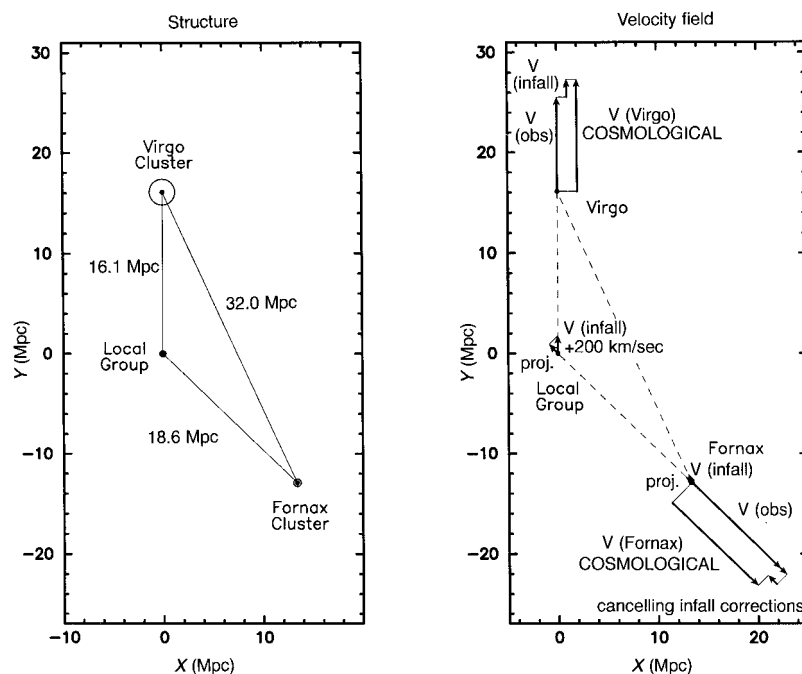


Figure 3 Relative geometry (left panel) and the corresponding velocity vectors (right panel) for the disposition and flow of the Fornax cluster and the Local Group with respect to the Virgo cluster. The circles plotted at the positions of the Virgo

and Fornax clusters have the same angular size as the circles minimally enclosing M100 and NGC1365 in the two panels of Fig. 1.

would increase from 70 to 74 km s⁻¹ Mpc⁻¹.

Given the highly clumped nature of the local Universe and the existence of large-scale streaming velocities, there is still an uncertainty remaining due to the total peculiar motion of the Fornax cluster with respect to the cosmic microwave background restframe. Observations of flows, and the determination of the absolute motion of the Milky Way with respect to the background radiation, suggest that line-of-sight velocities ~ 300 km s⁻¹ are not uncommon²⁰. The absolute motion of Fornax with respect to the Local Group then becomes the largest outstanding systematic uncertainty at this point in our error analysis: a 300 km s⁻¹ flow velocity for Fornax would give a systematic error in the Hubble constant of $\pm 20\%$.

We now step back and investigate the Hubble flow between us and Fornax, derived from galaxies and groups of galaxies inside 20 Mpc, each having Cepheid-based distances and expansion velocities and individually corrected for a Virgocentric flow model^{21,22}. Figure 4 presents these results in graphical form. Individual values of H_0 range from 61 to 99 km s⁻¹ Mpc⁻¹. An average of these independent determinations, including Virgo and Fornax, gives $H_0 = 73(\pm 4)[\pm 17]$ km s⁻¹ Mpc⁻¹. This determination, as before, uses a Virgocentric flow model with a $1/R_{\text{Virgo}}$ infall velocity fall-off, scaled to a Local Group infall velocity of +200 km s⁻¹, determined *ab initio* by minimizing the velocity residuals for the galaxies with Cepheid-based distances. The foregoing determination of H_0 is again predicated on the assumption that the inflow-corrected velocities of both Fornax and Virgo are not further perturbed by other mass concentrations or large-scale flows, such that the 25,000 Mpc³ volume of space delineated by them is at rest with respect to the distant galaxy frame. An estimated flow correction error of ± 300 km s⁻¹ dominates the systematic error budget. Although a full discussion of large-scale flows in general and the

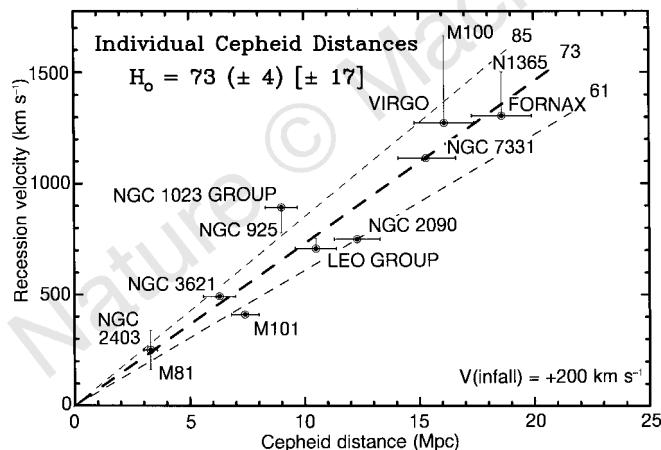


Figure 4 The velocity-distance relation for local galaxies having Cepheid-based distances with well defined velocities. Circled dots mark the velocities and distances of the parent groups or clusters. The one-sided "error" bars with galaxy names attached mark the velocities associated with the individual galaxies having direct Cepheid distances. The thick broken line represents a fit to the data giving $H_0 = 73$ km s⁻¹ Mpc⁻¹; the observed 1σ (random) scatter about the mean is ± 12 km s⁻¹ Mpc⁻¹, and is shown by the thin diverging broken lines. All velocities have been corrected for a Virgocentric flow, as discussed in the text. Galaxies associated with the Local Group and those in very nearby but widely dispersed groups (such as NGC5253 in the Centaurus group), as well as the individual galaxies (with Cepheid distances) along the line of sight to the Virgo cluster have been largely excluded from this analysis. Until large numbers of Cepheid distances become available for the other members of the nearest groups, it is impossible to associate a meaningful average distance with the averaged group velocity. As for Virgo, the median Cepheid distance derived from five galaxies along the line of sight (NGC4321, NGC4496A, NGC4536, NGC4571, NGC4639) is only 3% different from the distance to M100, which is adopted here for the Virgo cluster as a whole.

more local effects of the Great Attractor (for example) are beyond the scope of this Letter, the results reported here are consistent with those based on a more extensive analysis of the greater flow field, explicitly including the effects of the Great Attractor, as undertaken by M.H. *et al.* (manuscript in preparation), but inconsistent with the 500 km s⁻¹ infall of the Local Group into the Great Attractor derived by others^{23,24}.

We note the relatively small residual scatter in Fig. 4. Removing in quadrature the 10% scatter introduced by the known distance uncertainties leaves an average velocity scatter of $< \pm 70$ km s⁻¹ about the Hubble line. That is, the random velocities of nearby galaxies, once corrected for the systematic influence of the Virgo cluster, are remarkably small. This is consistent with earlier discussions³⁻⁵ that call to attention the paucity of nearby blue-shifted galaxies outside the Local Group.

Further implications of having a Cepheid distance to the Fornax cluster, and a comparison of the global value of the Hubble constant within the local expansion rate derived here, can be found in ref. 2. But we note that the estimates of H_0 presented here (based on nearby galaxies out to and including the Fornax cluster) are in close agreement with the value of $H_0 = 72$ km s⁻¹ Mpc⁻¹ obtained from a variety of methods that operate at much greater distances. These methods^{9,25,26} include the calibration of type Ia and type II supernovae, and the Tully-Fisher relation. □

Received 18 March; accepted 1 June 1998.

1. Freedman, W. L. *et al.* Distance to the Virgo cluster galaxy M100 from Hubble Space Telescope observations of Cepheids. *Nature* **371**, 757-762 (1994).
2. Madore, B. F. *et al.* The HST Key Project on the extragalactic distance scale. XV. Implications of a Cepheid distance to the Fornax cluster. *Astrophys. J.* (submitted).
3. Sandage, A. R. & Tammann, G. A. Steps toward the Hubble constant. V. The Hubble constant from nearby galaxies and the regularity of the local velocity field. *Astrophys. J.* **196**, 313-328 (1975).
4. Sandage, A. R., Tammann, G. A. & Hardy, E. Limits on the local deviation of the Universe from a homogeneous model. *Astrophys. J.* **172**, 253-263 (1972).
5. Fisher, J. R. & Tully, R. B. Neutral hydrogen observations of DDO dwarf galaxies. *Astron. Astrophys.* **44**, 151-171 (1975).
6. de Vaucouleurs, G. in *Stars and Stellar Systems* Vol. 9 (eds Sandage, A. R., Sandage, M. & Kristian, J.) 557-596 (Univ. Chicago Press, 1975).
7. Ferguson, H. C. & Sandage, A. R. Population studies in groups and clusters of galaxies. I. The luminosity function of galaxies in the Fornax cluster. *Astron. J.* **96**, 1520-1533 (1988).
8. Kennicutt, R. C., Freedman, W. L. & Mould, J. R. Measuring the Hubble constant with the Hubble Space Telescope. *Astron. J.* **110**, 1476-1491 (1995).
9. Freedman, W. L., Mould, J. R., Kennicutt, R. C. & Madore, B. F. *The Hubble Space Telescope Key Project to Measure the Hubble Constant* (ed. Sato, K.) (IAU Symp. 183, Reidel, in the press).
10. Ferguson, H. C. Population studies in groups and clusters of galaxies. II. A catalog of galaxies in the central 3.5 degrees of the Fornax cluster. *Astron. J.* **98**, 367-418 (1989).
11. Schroder, A. *UBVR Photometry of Spiral Galaxies in the Virgo and Fornax Clusters* Thesis, Univ. Basel (1995).
12. Han, M. & Mould, J. R. The velocity field in the local supercluster. *Astrophys. J.* **360**, 448-464 (1990).
13. Bureau, M., Mould, J. R. & Staveley-Smith, L. A new I-band Tully-Fisher relation for the Fornax cluster: implication for the Fornax distance and local supercluster velocity field. *Astrophys. J.* **463**, 60-68 (1996).
14. Silbermann, N. *et al.* The HST Key Project on the extragalactic distance scale. XIV. Cepheids in NGC 1365. *Astrophys. J.* (submitted).
15. Madore, B. F. & Freedman, W. L. The Cepheid distance scale. *Publ. Astron. Soc. Pacif.* **103**, 933-957 (1991).
16. Feast, M. W. & Catchpole, R. M. The Cepheid period-luminosity zero-point from Hipparcos trigonometrical parallaxes. *Mon. Not. R. Astron. Soc.* **286**, L1-L5 (1997).
17. Madore, B. F. & Freedman, W. L. Hipparcos parallaxes and the Cepheid distance scale. *Astrophys. J.* **492**, 110-115 (1998).
18. Sandage, A. R. & Tammann, G. A. Confirmation of previous ground-based Cepheid P-L zero-points using Hipparcos trigonometric parallaxes. *Mon. Not. R. Astron. Soc.* **293**, L23-L26 (1998).
19. Schechter, P. Mass-to-light ratios for elliptical galaxies. *Astron. J.* **85**, 801-811 (1980).
20. Coles, P. & Lucchin, F. *Cosmology* 399-400 (Wiley, New York, 1995).
21. Kraan-Korteweg, R. A catalog of 2810 nearby galaxies: the effect of the Virgocentric flow model on their observed velocities. *Astron. Astrophys. Suppl.* **66**, 255-279 (1986).
22. Aaronson, M., Huchra, J., Mould, J., Schechter, P. & Tully, R. B. The velocity field in the local supercluster. *Astrophys. J.* **258**, 64-76 (1980).
23. Faber, S. M. & Burstein, D. in *Large-Scale Motions in the Universe* (eds Rubin, V. C. & Coyne, G. V.) 116-167 (Princeton Univ. Press, 1988).
24. Marinoni, C., Monaco, P., Giuricin, G. & Constantini, B. Galaxy distances in the nearby universe: corrections for peculiar motions. *Mon. Not. R. Astron. Soc.* (submitted).
25. Reiss, A., Press, W. & Kirshner, R. Using type Ia supernova light curve shapes to measure the Hubble constant. *Astrophys. J.* **438**, L17-L20 (1995).
26. Giovanelli, R. *et al.* The Tully-Fisher relation and H_0 . *Astrophys. J.* **477**, L1-L4 (1997).

Acknowledgements. This work was supported by NASA and the NSF, and made extensive use of the NASA/IPAC Extragalactic Database (NED). Observations are based on data obtained using the Hubble Space Telescope which is operated by the Space Telescope Science Institute under contract from the Association of Universities for Research in Astronomy. L.F. acknowledges support by NASA through a Hubble Fellowship grant awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA.

Correspondence and requests for materials should be addressed to B.F.M. (e-mail: barry@ipac.caltech.edu).

Most smooth closed space curves contain approximate solutions of the n -body problem

Gregory Buck

Department of Mathematics, Saint Anselm College, Manchester, New Hampshire 03102, USA

The determination of the exact trajectories of mutually interacting masses (the n -body problem^{1,2}) is apparently intractable for $n \geq 3$, when the generic solutions become chaotic. A few special solutions are known, which require the masses to be in certain initial positions; these are known as 'central configurations' (refs 1–6) (an example is the equilateral triangle formed by the Sun, Jupiter and Trojan asteroids). The configurations are usually found by symmetry arguments. Here I report a generalization of the central-configuration approach which leads to large continuous families of approximate solutions. I consider the uniform motion of equidistributed masses on closed space curves, in the limit when the number of particles tends to infinity. In this situation, the gravitational force on each particle is proportional to the local curvature, and may be calculated using an integral closely related to the Biot–Savart integral. Approximate solutions are possible for certain (constant) values of the particle speed, determined by equating this integral to the mass times the centrifugal acceleration. Most smooth, closed space curves contain such approximate solutions, because only the local curvature is involved. Moreover, the theory also holds for sets of closed curves, allowing approximate solutions for knotted and linked configurations.

Central configurations^{1–6} are usually thought of in one of two ways. With appropriate initial velocities, a planar central configuration gives homographic solutions—in which the shape of the initial configuration is preserved, but not necessarily the size (a special case is rigid rotation, where the centrifugal forces exactly balance the attraction due to gravity and size is preserved as well). In our Solar System, the Sun, Jupiter and the Trojan asteroids form an equilateral triangle homographic solution. Alternatively, if a central configuration is allowed to run from rest, then the solution is that the configuration collapses to the centre of mass, while keeping its shape.

Here I generalize the idea of a central configuration, searching for conditions such that a set of masses would move along a given space curve. Let $\mathbf{x}_i$ denote the position vector for the i th mass, and let $\ddot{\mathbf{x}}_i$ denote the acceleration vector of the i th mass. Consider the regular n -gon of equal masses, which is known to be a central configuration for all n . If the velocities are of the correct magnitude and tangent to the circle containing the n -gon, then a relative equilibrium solution is given; the masses move along the circle with uniform speed. Let κ denote the curvature vector—the vector in the direction of the normal with length equal to the curvature. Then in this system we have $\kappa = (1/v^2)\ddot{\mathbf{x}}_i$ for all masses in all positions $\mathbf{x}_i(t)$ along the circle containing the masses, because if a mass travels along a given curve with uniform speed v , $\ddot{\mathbf{x}}_i = \kappa v^2$.

In general, if there were n masses distributed along a closed space curve, moving along the curve with uniform speed v , and the condition $\kappa = (1/v^2)\ddot{\mathbf{x}}_i$ held for all masses in all positions $\mathbf{x}_i(t)$ (all translations of the masses along the curve), then the space curve would contain all the trajectories of the system, and we would have a solution to the n -body problem. In this sense, the equations $\ddot{\mathbf{x}}_i = \kappa v^2$ can be thought of as a generalization of the central configuration equations $\ddot{\mathbf{x}}_i = \kappa \mathbf{x}_i$.

I now consider the likelihood of finding such curves. The perhaps surprising result is that, in the limit as n (the number of masses) tends to ∞ , most smooth closed space curves are such curves. By 'most' we mean curves with bounded curvature, a finite number of inflexion points and without self-intersections. These are generic conditions for smooth space curves.

Consider a sub-arc of such a curve, and let n , the number of equal masses distributed along the curve, be large enough that the change in curvature is small between successive masses (Fig. 1). Choose a mass m_i . The components of $\ddot{\mathbf{x}}_i$ come in pairs, taking the effect from m_{i+1} together with the force from m_{i-1} , m_{i+2} with m_{i-2} , and so on. As the masses are equidistributed along the curve, and the curve is nearly 'flat' between successive masses, this pairing cancels the tangential component of the force from the nearest neighbours. To first approximation we are left with the component of the force in the direction of κ .

Next I consider the limit of a distribution along the curve, keeping the total mass constant (say 1), and letting there be n masses of mass $1/n$, and letting $n \rightarrow \infty$. Suppressing for the moment G , the gravitational constant, and ρ , the line density (mass per unit length), which both enter as scalars, the computation of $\ddot{\mathbf{x}}$ becomes an integral, $\int_c (\mathbf{x} - \mathbf{y})/|\mathbf{x} - \mathbf{y}|^3$, where $\mathbf{x}$ is a particular point on the curve c , and $\mathbf{y}$ varies along the curve. The cancellation of tangent forces mentioned above means that the integral, near the point $\mathbf{x}$, can be approximated by $(\kappa/|\kappa|) \int (\cos \alpha)/|\mathbf{x} - \mathbf{y}|^2$, where α is the angle between the vector $\mathbf{x} - \mathbf{y}$ and κ , the curvature vector at $\mathbf{x}$. This integral is divergent as $\mathbf{y}$ approaches $\mathbf{x}$. The $\cos \alpha$ cancels one power of $\mathbf{x} - \mathbf{y}$ at the singularity, but the curvature gives a coefficient to this cancellation. The remaining inverse power gives us a logarithmic singularity at $\mathbf{x} = \mathbf{y}$ (Fig. 1). Let R be the local radius of curvature, and δ the small particle size ($1/n$). Then in the limit as $n \rightarrow \infty$ we have $\ddot{\mathbf{x}}_i \rightarrow G\rho \log(R/\delta)\kappa$. Here I used the case of equal masses for ease of exposition. The analysis requires only the equidistribution of mass along the curve.

Let the particles move along the curve with uniform speed v given by $v^2 = G\rho \log(R/\delta)$. (R weakly varies over c , but as $n \rightarrow \infty$ it tends to

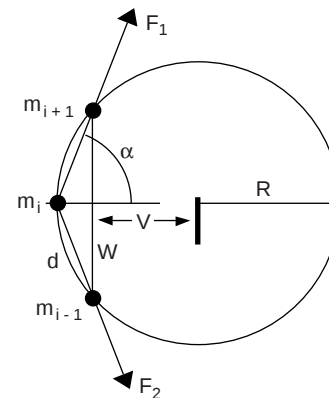


Figure 1 The gravitational force vectors $\mathbf{F}_1$ and $\mathbf{F}_2$ result from the effect of m_{i+1} and m_{i-1} on m_i . As m_{i+1} and m_{i-1} approach m_i symmetrically along the circle, the components of $\mathbf{F}_1$ and $\mathbf{F}_2$ tangential to the curve (here the circle) cancel, leaving only the contribution in the direction of the curvature vector. Passing to the continuous case, it is necessary to analyse $\int (\cos \alpha)/d^2$. But using $d^2 = w^2 + (R - v)^2$, we have $\cos \alpha/d^2 = (1/d)[(R - v)/d^2] = (1/d)[(R - v)/(2R^2 - 2Rv)] = (1/d)(1/2R)$, so the integral becomes $(1/2)\kappa \int (1/d)$, which is the logarithmic singularity times the curvature vector. The analysis for the general case is given by considering the osculating circle at the given point on the curve. This analysis is similar to the calculation carried out in fluid dynamics in the analysis of the dynamics of vortex filaments. In that case, the force is in the direction of the binormal of the underlying curve (the vortex filament), as opposed to the normal, and the logarithmic singularity is regularized, giving the local induction approximation: the vortex filament's self induction is given by speed that is the norm of the curvature, in the direction of the binormal vector^{7,8}.

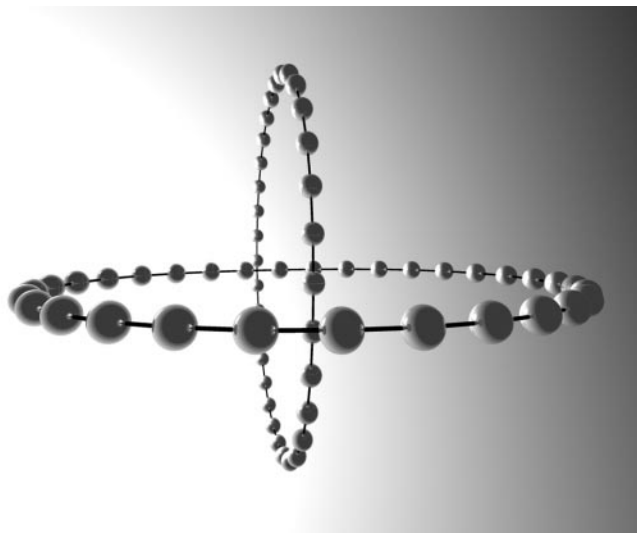


Figure 2 Schematic representation of a simple link approximate solution.

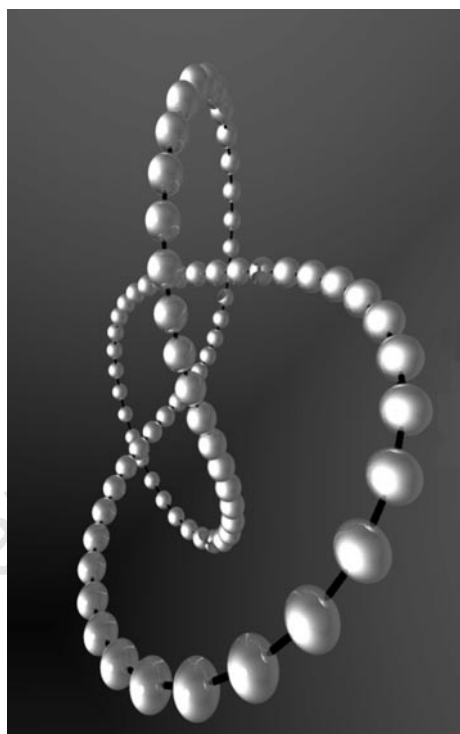


Figure 3 Schematic representation of a knotted approximate solution. One way to measure the rate of convergence (in n) to the infinite speed solution for a particular curve is by subtracting off the logarithmic singularity at each point. Consider the double integral $E(c) = \iint [(\mathbf{x} - \mathbf{y})/|\mathbf{x} - \mathbf{y}|^3] \cdot [(\mathbf{x} - \mathbf{S})/|\mathbf{x} - \mathbf{S}|^3]$, where the interior integral is computed by holding $\mathbf{x}$ fixed and letting $\mathbf{y}$ vary over the curve c , and letting $\mathbf{S}$ vary over the osculating circle to c at $\mathbf{x}$. For the exterior integral, let $\mathbf{x}$ vary over c . The integral is finite for the curves which satisfy our conditions. $E(\text{circle}) = 0$, as the osculating circle for the circle is the circle itself. A higher E value indicates that more masses may be required for a better approximate solution. For example, if the curve self-intersects, $E(c) = \infty$, and near self-intersections have high $E(c)$. $E(c)$ is closely related to 'energy' functions for knots and links that have recently been studied⁹⁻¹². Such functions have been found to be remarkably effective at simplifying complex tangles (R. Scharein, Knot-Plot, program available at <http://www.cs.ubc.ca/nest/imager/contributions/scharein/KnotPlot.html>; K. Brakke, Evolver, program available at <http://www.geom.umn.edu>), and have found applications in biochemistry^{13,14}.

a uniform value, so for uniform motion a representative constant should be chosen). Then Newton's equations of motion are nearly satisfied for all times.

A physical interpretation is to think of the bodies as restricted to the space curve, like beads on a wire, and ask how much force is used to keep the beads on the wire as they traverse it (at speed v). In a solution, such as a relative equilibrium, no force at all is required. In an approximate solution this force should be small. In this spirit I offer a definition of periodic approximate solution—other definitions are certainly possible. In a solution to the n -body problem, Newton's equations of motion

$$m_i \ddot{\mathbf{x}}_i = \sum_{j=1, j \neq i}^n \frac{Gm_i m_j (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_i - \mathbf{x}_j|^3}$$

hold for all masses. In an approximate solution, which is a set of trajectories $\mathbf{X}(t) = \mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)$, the equations nearly hold. It is necessary to define 'nearly'. Let

$$E_i = \left| m_i \ddot{\mathbf{x}}_i - \sum_{j=1, j \neq i}^n \frac{Gm_i m_j (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_i - \mathbf{x}_j|^3} \right|$$

For periodic trajectories, let E_r denote the maximum over the n masses of $E_i \omega$, where ω is the period and E_i is the maximum of E_i over $0 \leq t \leq \omega$. If E_r is small, there is a sort of uniform bound on the 'error' along each trajectory and I will call $\mathbf{X}(t)$ an approximate solution. For the trajectories along closed curves described above $E_r \rightarrow 0$ as $n \rightarrow \infty$. Simpler sets of curves, such as the Hopf link (Fig. 2), require fewer masses to make E_r small. Measures of complexity of curves may give us estimates of the number of masses required (Fig. 3).

Questions remain regarding the exact relationship of these approximate solutions to n -body initial value problems. In particular, consider the initial value problem of many masses equidistributed along a closed curve, with initial velocities which are tangent to the curve and have constant magnitude $v = (G\rho \log(R/\delta))^{1/2}$. The existence of the approximate solutions would seem to give an idea of the behaviour of the initial value problem for some time interval—the masses nearly follow the underlying curve: the question is how long a time interval.

A long-standing open question in celestial mechanics is whether there exist continuous families of central configurations for a given set of masses. We can construct approximate collapse solutions consisting of masses distributed along circles about the centre of mass, where the density of the distribution on a given circle of radius R is R^2 (so we have $\ddot{\mathbf{x}}_i = k\mathbf{x}_i$ for masses on each circle). As the local contribution along the curve dominates, the orientation of the

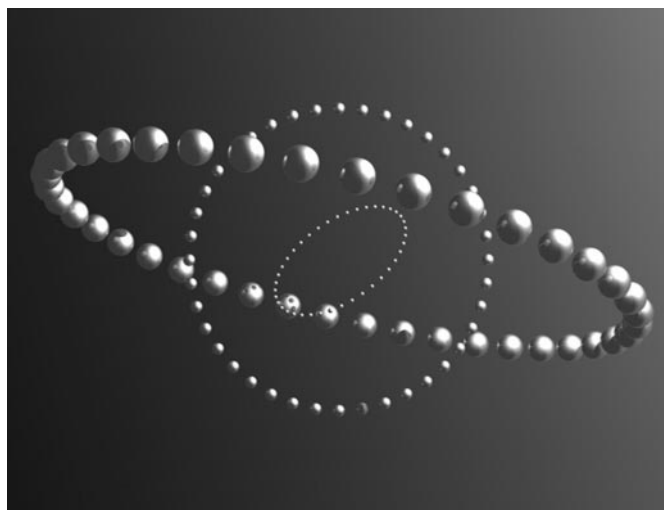


Figure 4 Schematic of approximate collapse solutions.

circles to one another makes no difference, and in the limit we have continuous families of collapse solutions (Fig. 4).

The analysis involved an integral of the same form as the Biot–Savart integral, a much studied integral which arises in several contexts in magneto-hydrodynamics. This connection invites further investigation. The results may be of use in particle systems where filamentary distributions tend to arise, such as cosmic strings. Also, because the acceleration of a filamentary equidistribution is approximated by the curvature, connections to curvature driven flows are possible. □

Received 20 January; accepted 6 July 1998.

1. Diacu, F. & Holmes, P. *Celestial Encounters* (Princeton Univ. Press, 1996).
2. Meyer, K. & Hall, G. R. *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem* (Springer, New York, 1992).
3. Saari, D. On the role and properties of n-body central configurations. *Celest. Mech.* **21**, 9–20 (1980).
4. Moeckel, R. On central configurations. *Math. Z.* **205**, 499–517 (1990).
5. Buck, G. On clustering in central configurations. *Proc. Am. Math. Soc.* **108**, 801–810 (1990).
6. Llibre, J. On the number of central configurations in the n-body problem. *Celest. Mech.* **50**, 89–96 (1991).
7. Ricca, R. L. Rediscovery of Da Rios equations. *Nature* **352**, 561–562 (1990).
8. Saffman, P. G. *Vortex Dynamics* (Cambridge Univ. Press, 1992).
9. Buck, G. Four-thirds power law for knots and links. *Nature* **392**, 238–239 (1998).
10. Buck, G. & Simon, J. Thickness and crossing number of knots. In *Topol. Appl.* (in the press).
11. Moffatt, H. K. The energy spectrum of knots and links. *Nature* **347**, 367–369 (1990).
12. O'Hara, J. Energy of a knot. *Topology* **30**, 241–247 (1991).
13. Simon, J. in *Proc. 1994 IMA Summer Program on Molecular Biology* (eds Mesirov, J. P., Schulten, K. & Sumners, D. W.) 39–58 (IMA Vol. 82, Springer, New York, 1996).
14. Stasiak, A. et al. Electrophoretic mobility of DNA knots. *Nature* **384**, 122 (1996).

Acknowledgements. This work was supported in part by the National Science Foundation.

Correspondence and requests for materials should be addressed to the author (e-mail: gbuck@anselm.edu).

Strong exciton–photon coupling in an organic semiconductor microcavity

D. G. Lidzey^{††}, D. D. C. Bradley^{††}, M. S. Skolnick^{*}, T. Virgili^{††}, S. Walker[‡] & D. M. Whittaker[§]

^{*} Department of Physics, ^{††} Centre for Molecular Materials, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, UK

[‡] Department of Electronic and Electrical Engineering, University of Sheffield, Sir Frederick Mappin Building, Mappin Street, Sheffield S1 3JD, UK

[§] Toshiba Cambridge Research Centre Ltd, 260 Cambridge Science Park, Milton Road, Cambridge CB4 4WE, UK

The modification and control of exciton–photon interactions in semiconductors is of both fundamental^{1–4} and practical interest, being of direct relevance to the design of improved light-emitting diodes, photodetectors and lasers^{5–7}. In a semiconductor microcavity, the confined electromagnetic field modifies the optical transitions of the material. Two distinct types of interaction are possible: weak and strong coupling^{1–4}. In the former perturbative regime, the spectral and spatial distribution of the emission is modified but exciton dynamics are little altered. In the latter case, however, mixing of exciton and photon states occurs leading to strongly modified dynamics. Both types of effect have been observed in planar microcavity structures in inorganic semiconductor quantum wells and bulk layers^{1–8}. But organic semiconductor microcavities have been studied only in the weak-coupling regime^{9–18}. Here we report an organic semiconductor microcavity that operates in the strong-coupling regime. We see characteristic mixing of the exciton and photon modes (anti-crossing), and a room-temperature vacuum Rabi splitting (an indicator of interaction strength) that is an order of magnitude larger than the previously reported highest values for inorganic semiconductors. Our results may lead to new structures and device concepts incorporating hybrid states of organic and inorganic excitons¹⁹, and suggest that polariton lasing^{20–22} may be possible.

Strong coupling manifests itself in anti-crossing of the coupled modes and the appearance, on resonance, of two equal-intensity transitions separated by the vacuum Rabi splitting. These new cavity polariton modes can be considered an admixture of the optical transitions of the material and the cavity photon modes, and have a radiative lifetime determined by the cavity photon lifetime. Strong coupling is evident when the interaction strength (Rabi splitting) exceeds the optical transition and cavity mode damping. Under these conditions, a periodic exchange of energy (Rabi oscillation) can occur between the coupled modes before the excitation decays. In the perturbative weak-coupling regime, the interaction strength is less than the damping. Anti-crossing is then no longer observed and whilst significant changes in the spectral and spatial distribution of transition probability do still occur, no dramatic changes in recombination dynamics are expected.

Up until now, strong coupling in a semiconductor microcavity has only been achieved using inorganic materials. A typical structure consists of a high finesse Fabry–Pérot resonator with distributed Bragg reflector mirrors and a series of quantum wells placed close to the confined field antinode. The largest Rabi splitting reported for a III–V (GaAs) based microcavity structure is²³ 9.4 meV. Larger values, namely 17.5 meV, have been reported for II–VI (ZnCdSe)-based microcavities⁸. The splitting depends on the exciton oscillator strength and the magnitude of the confined field, and is maximized for matched cavity and optical transition linewidths^{1–4,24}. The enhanced Rabi splitting for II–VI semiconductors is principally due to their larger oscillator strengths.

Organic semiconductors have recently attracted much interest for applications in electroluminescent displays and as gain media. Microcavity structures^{9–18} have been fabricated to allow control of the emission energy, linewidth, intensity and directionality for light-emitting diodes, as prototypical colour displays and as laser resonators. Large oscillator strengths are a characteristic feature of these materials. Strong coupling has not, however, been reported. The large exciton linewidths²⁵ that result from inhomogeneous broadening and the presence of a vibronic progression make strong coupling difficult to observe. This situation is not, however, universally true and we have identified several candidate materials that have sufficiently narrow exciton linewidths that strong-coupling might be expected to be seen in a typical microcavity structure.

We report experiments that use tetra-(2,6-*t*-butyl)phenol-porphyrin zinc (4TBPPZn) as the organic semiconductor. This molecule has

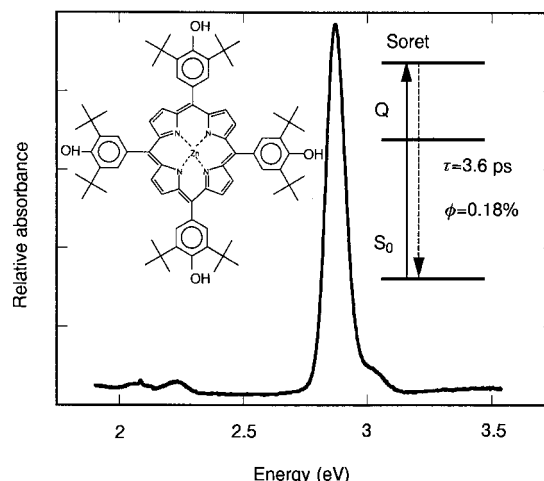


Figure 1 The absorption spectrum of a blend film of 4TBPPZn in polystyrene. The insets show the chemical structure and electronic energy-level scheme^{26,27} for 4TBPPZn. The Soret band peaks at 2.88 eV and the Q-band absorbance is seen as a weak double peak structure between 2.0 and 2.26 eV. The Soret-band exciton decay time is expected to be $\tau = 3.6$ ps and the emission yield $\phi = 0.18\%$, due to rapid relaxation to the Q-band exciton²⁷.

the desired intense narrow exciton transition in the form of the so-called 'Soret band' absorption at 2.88 eV. 4TBPPZn is one example of a large family of porphyrin macrocycles whose photophysical and photochemical properties have been extensively studied^{26,27}. The chemical structure, optical absorption spectrum and electronic energy-level scheme for 4TBPPZn are shown in Fig. 1.

Figure 2 shows the general structure of the microcavities used in our experiments. The room-temperature reflectivity of a typical microcavity measured at normal incidence to the glass substrate is shown in Fig. 3 trace a. As the cavity mode is tuned closer to the exciton energy, strong coupling is observed (Fig. 3 trace b) and on resonance the expected pair of equal intensity Rabi-split transitions are seen (Fig. 3 trace c). For larger angles, beyond resonance (Fig. 3 trace d) the coupling reduces, again in line with expectation. The energies of the two transitions are plotted in Fig. 4a as a function of the angle of incidence. The cavity photon and 4TBPPZn exciton modes exhibit a very clear anti-crossing behaviour and the Rabi-

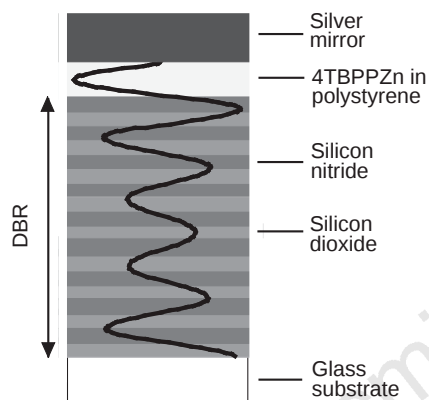


Figure 2 The general structure of the $\lambda/2$ microcavities used in this work. The nine alternating pairs of silicon nitride (refractive index $n = 1.95$) and silicon dioxide ($n = 1.45$) $\lambda/4$ layers of the distributed Bragg reflector (DBR) mirror were grown by plasma-enhanced chemical vapour deposition. At normal incidence these mirrors had peak reflectivities of 0.992 at 2.76 eV. A blend of polystyrene and 4TBPPZn was spin-coated from toluene solution on top of the DBR mirror to give films of thickness $L \approx 100$ nm. The second mirror of the microcavity was produced by evaporation of silver onto this film. Also shown superimposed on the structure is a transfer matrix reflectivity simulation of the confined optical field at the energy of the cavity mode. The confined field possesses an antinode in the centre of the organic layer to maximize exciton-photon coupling. Typical Q values for these structures were ≈ 125 . The peak absorbance, αL , where α is the absorption coefficient of the 4TBPPZn Soret band of the blend film and L the cavity length, was varied in the range 0.23 to 1.05 by changing the weight fraction of 4TBPPZn in the blend. This readily executed variation of absorbance allows modification of the exciton-photon interaction strength (see Fig. 4b and text) without changing the cavity length (an attractive feature of the use of a blend system). The polystyrene matrix also ensures that intermolecular interactions between 4TBPPZn molecules, which can cause line broadening, are minimized. The resultant films had a Soret band full-width at half-maximum (FWHM) linewidth (which increased slightly with increasing 4TBPPZn concentration) of 90 ± 5 meV (≈ 15 nm) at room temperature. With a silver top mirror, the cavity mode FWHM linewidths for the microcavities studied were found to lie in the range 18–25 meV (≈ 3.5 nm). The energy of the cavity photon mode is a strong function of the angle of incidence relative to the surface normal: as for a standard Fabry-Pérot etalon, the mode energy varies as $1/\cos\theta_{\text{int}}$ where θ_{int} is the internal angle within the cavity. For increasing angles, the cavity mode steadily shifts to higher energy. This phenomenon is a well known problem for display applications of microcavities¹⁵ but here it allows tuning of the interaction between the exciton and photon modes: the exciton transition energy is angle-independent, and hence changing the angle of incidence allows one to adjust the relative separation of the two modes. Microcavities were fabricated such that the cavity mode was at lower energy than the exciton transition at normal incidence, and become resonant with it at $\sim 40^\circ$.

splitting energy, $\hbar\Omega \approx 110$ meV (where Ω is the angular frequency of Rabi oscillation) between the two transitions at the 40° resonance angle is exceptionally large.

The Rabi splitting can be varied by adjusting the composition of the blend films to control the 4TBPPZn exciton absorbance. The splitting is expected, to first approximation, to vary as the square root of the absorbance²⁴. Figure 4b shows the $\hbar\Omega$ values of a series of microcavities plotted as a function of $(\alpha L)^{1/2}$ where α is the absorption coefficient of the 4TBPPZn Soret band of the blend film and L is the cavity length: the observed proportionality is in good agreement with expectation. The largest room-temperature Rabi splitting that we have measured, $\hbar\Omega \approx 160$ meV, is an order of magnitude larger than the largest splittings observed for II–VI quantum-well microcavities⁸ and more than 30 times the typical splittings observed for III–V quantum-well microcavities at room temperature^{1–4}. Such a significant difference is principally attributable to a very large exciton oscillator strength. For comparison with work on III–V quantum wells, we have calculated (using a transfer matrix reflectivity model) the effective 4TBPPZn exciton oscillator strength per unit area, f , for the blend film with $\hbar\Omega \approx 160$ meV: we find a value $f \approx 2 \times 10^{15} \text{ cm}^{-2}$. This compares with a value of $f \approx 4.2 \times 10^{12} \text{ cm}^{-2}$ per quantum well for an InGaAs structure⁴. Using the relation⁴ $\hbar\Omega \propto (f/n_c^2 L_{\text{eff}})^{1/2}$, where n_c is the refractive index in the cavity and L_{eff} its effective length (taking account of penetration of the optical field into the distributed Bragg reflector), we expect $\hbar\Omega$ to be a factor of ~ 40 larger than for a microcavity with three InGaAs quantum wells. The experimental Rabi splitting is 32 times larger than the $\hbar\Omega \approx 5$ meV splitting⁴ for the InGaAs quantum wells at 5 K. The enhanced splitting arises both from an increased effective oscillator strength (~ 10 times enhancement) and from better confinement of the optical field due to the low dielectric constant of the blend film and the short effective cavity length (~ 3 times enhancement for the measured $n_c = 1.63$ and calculated $L_{\text{eff}} = 389$ nm). We note that further enhancement of the Rabi splitting might be expected if the 4TBPPZn exciton linewidth could be reduced closer to that of the photon mode, bringing the result into even better agreement.

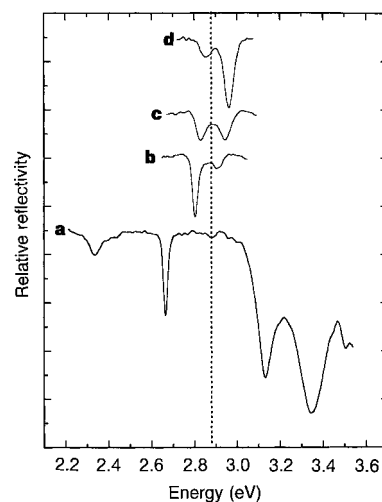


Figure 3 Reflectivity spectra measured at room temperature as the cavity photon mode is angle-tuned through the 4TBPPZn exciton transition. (See Fig. 2 legend for details of angle-tuning.) In spectrum a, the photon mode is far from resonance. The high reflectivity stop-band of the resonator extends from 2.4 to 3 eV. The cavity photon mode shows up within the stop-band as a sharp dip in the reflectivity with a minimum at 2.68 eV. The 4TBPPZn exciton absorption at 2.88 eV can only just be detected (marked by the dashed vertical line) due to the low transmission (4%) of the dielectric mirror at this energy. In spectrum b, the photon mode approaches resonance, in spectrum c it is on-resonance, and in spectrum d it is moving away from resonance again. We note that the spectra are displaced vertically for clarity.

The large Rabi splitting may make these structures of interest for fundamental studies of coherent dynamics in a strongly coupled exciton–photon system. Rabi oscillations (beats between the two coupled modes) are observed when the system is excited on-resonance with a coherent optical pulse of bandwidth greater than the Rabi splitting. In the absence of faster dephasing processes, the oscillations with period τ_{Rabi} are damped in a time governed by the cavity photon lifetime τ_{photon} which is much shorter than the decay time from the exciton state. For our organic microcavities, one expects $\tau_{\text{photon}} \approx 30$ fs; with a splitting of 160 meV, τ_{Rabi} is only 26 fs, so that many periods of Rabi oscillation can be expected. Another area for study concerns the high excitation density properties of coupled exciton–photon systems. The large exciton binding energy of organic semiconductors (~ 0.25 – 0.5 eV; refs 28, 29) suggests that the strong-coupling regime should survive to much higher excitation density than for III–V semiconductor microcavities^{20,21}, and thus that polariton lasing might be expected, as reported for II–VI semiconductors²².

The typical radiative lifetime of organic semiconductor excitons is ~ 1 ns (ref. 30), but non-radiative decay processes associated with coupling to vibrational modes (energies ~ 150 – 200 meV) often lead to strongly reduced decay times and emission efficiencies: for tetraphenylporphyrin zinc²⁷, the Soret-band decay time is 3.6 ps and the emission yield only 0.18%. The decay time for the polariton states (estimated to be ~ 60 fs for our structures on-resonance) is

clearly much shorter than this, and is in fact comparable to the period of the vibrational modes through which non-radiative decay occurs. This may mean that non-radiative decay channels that normally compete with exciton radiative decay could be circumvented, and hence that the emission yield could be enhanced. Measurements addressing this point are in progress.

The demonstration of strong coupling involving organic semiconductor excitons also opens the way to structures in which new hybrid states tailored to control emission properties could be produced. Recent theoretical interest has focused on the possibility of coupling organic and inorganic excitons via a cavity photon¹⁹. In such structures, it is proposed that excitons electrically created in an inorganic quantum well would rapidly relax into more highly emissive polariton states derived largely from the organic excitons. If the local density of these states is not prohibitively low, such coupling should help to increase radiative rates by reducing the rate-limiting slow relaxation associated with acoustic phonon emission. Coupling of exciton states in two or more organic semiconductors is also possible, and could lead to new concepts for emissive devices.

We note that the observation of strong coupling is not confined to porphyrin molecules like 4TBPPZn. We find similar behaviour for cyanine dyes (D.G.L., T.V., D.D.C.B., M.S.S., S.W. and D.M.W., manuscript in preparation). □

Received 31 March; accepted 19 June 1998.

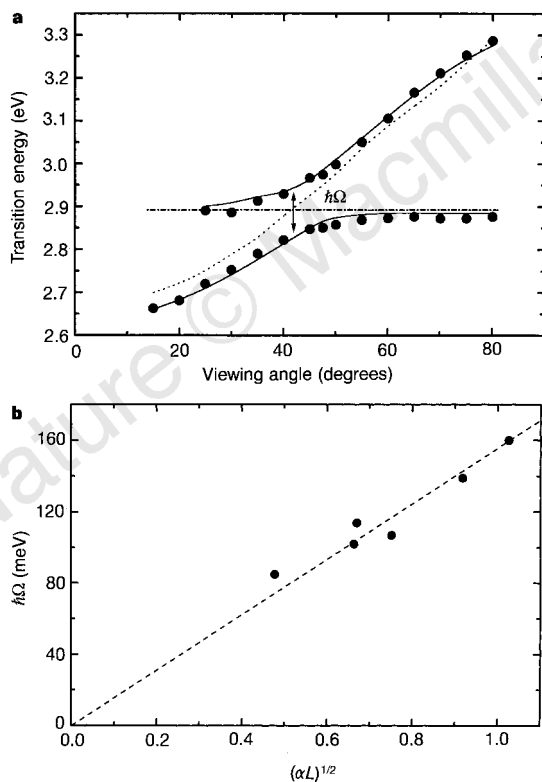


Figure 4 Evaluation of the Rabi splitting energy and demonstration of its dependence on exciton oscillator strength. **a**, Angular dispersion of the coupled modes in a typical microcavity. The horizontal dash-dotted line at 2.88 eV is the 4TBPPZn exciton transition energy in the polystyrene blend film. This energy is angle-independent, and so is expected to give the resonance energy for the coupled exciton–photon system. The dotted line shows the measured angular dispersion of the photon mode for a control microcavity that contained polystyrene but no 4TBPPZn ($\alpha = 0$). The solid line is a fit to the data using a transfer matrix reflectivity model from which the oscillator strength can be deduced. The on-resonance Rabi-splitting energy, $\hbar\Omega$, between the two transitions is marked. **b**, Variation of the Rabi splitting as a function of $(\alpha L)^{1/2}$, where αL is the absorbance of the blend film. The dashed line is a guide to the eye.

1. Weisbuch, C. *et al.* Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Phys. Rev. Lett.* **69**, 3314–3317 (1992).
2. Houdré, R. *et al.* Room-temperature cavity polaritons in a semiconductor microcavity. *Phys. Rev. B* **49**, 16761–16764 (1994).
3. Burstein, E. & Weisbuch, C. (eds) *Confined Electrons and Photons* (Plenum, New York, 1995).
4. Skolnick, M. S., Fisher, T. A. & Whittaker, D. M. Strong coupling phenomena in quantum microcavity structures. *Semiconductor Science & Technology* (in the press).
5. Yamamoto, Y. *et al.* in *Coherence, Amplification, and Quantum Effects in Semiconductor Lasers* (ed. Yamamoto, Y.) 561–615 (Wiley, New York, 1991).
6. Unlü, M. S. & Strite, S. Resonant cavity enhanced photonic devices. *J. Appl. Phys.* **78**, 607–638 (1995).
7. Sale, T. E. *Vertical Cavity Surface Emitting Lasers* (Wiley, New York, 1995).
8. Kelkar, P. *et al.* Excitons in a II–VI semiconductor microcavity in the strong-coupling regime. *Phys. Rev. B* **52**, R5491–R5494 (1995).
9. Nakayama, T. & Kakuta, A. Organic luminescent devices with a multiplex cavity structure. *Jpn. J. Appl. Phys.* **34**, L1234–L1236 (1995).
10. Tsutsui, T. *et al.* Sharply directed emission in organic electroluminescent diodes with an optical-microcavity structure. *Appl. Phys. Lett.* **65**, 1868–1870 (1994).
11. Dodabalapur, A. *et al.* Microcavity effects in organic semiconductors. *Appl. Phys. Lett.* **64**, 2486–2488 (1994).
12. Jordan, R. H. *et al.* Efficiency enhancement of microcavity organic light emitting diodes. *Appl. Phys. Lett.* **69**, 1997–1999 (1996).
13. Lidzey, D. G. *et al.* Control of photoluminescence emission from a conjugated polymer using an optimised microcavity structure. *Chem. Phys. Lett.* **263**, 655–660 (1996).
14. Lidzey, D. G. *et al.* Mapping the confined field in a microcavity via the emission from a conjugated polymer. *Appl. Phys. Lett.* **71**, 744–746 (1997).
15. Lidzey, D. G. *et al.* Pixelated multicolour microcavity displays. *Spec. Topics Quantum Electron.* **4**, 113–118 (1998).
16. Grüner, J. *et al.* Optical-mode structure in a single-layer polymer microcavity. *Synth. Met.* **76**, 137–140 (1995).
17. Tessler, N. *et al.* Lasing from conjugated-polymer microcavities. *Nature* **382**, 695–697 (1996).
18. Diaz-Garcia, M. A. *et al.* Plastic lasers: comparison of gain narrowing with a soluble semiconducting polymer in waveguides and microcavities. *Appl. Phys. Lett.* **70**, 3191–3193 (1997).
19. Agranovich, V. *et al.* Organic and inorganic quantum wells in a microcavity: Frenkel–Wannier–Mott excitons hybridization and energy transformation. *Solid State Commun.* **102**, 631–636 (1997).
20. Kira, M. *et al.* Quantum theory of nonlinear semiconductor microcavity luminescence explaining “boser” experiments. *Phys. Rev. Lett.* **79**, 5170–5175 (1997).
21. Jahnke, F. *et al.* Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Phys. Rev. Lett.* **77**, 5257–5260 (1996).
22. Kelkar, P. V. *et al.* Stimulated emission, gain, and coherent oscillations in II–VI semiconductor microcavities. *Phys. Rev. B* **56**, 7564–7573 (1997).
23. Nelson, T. R. *et al.* Room-temperature normal-mode coupling in a semiconductor microcavity utilizing native-oxide Al Al/GaAs mirrors. *Appl. Phys. Lett.* **69**, 3031–3033 (1996).
24. Norris, T. B. in *Confined Electrons and Photons* (eds Burstein, E. & Weisbuch, C.) 503–521 (Plenum, New York, 1995).
25. Grice, A. W. *et al.* A blue emitting triazole based conjugated polymer. *Adv. Mater.* **9**, 1174–1178 (1997).
26. Dolphin, D. (ed.) *The Porphyrins* (Academic, New York, 1978).
27. Kalyanasundaram, K. *Photochemistry of Polypyridine and Porphyrin Complexes* (Academic, New York, 1992).
28. Marks, R. N. The photovoltaic response in poly(*p*-phenylene vinylene) thin film devices. *J. Phys. Cond. Matt.* **6**, 1379–1394 (1994).
29. Alvarado, S. *et al.* Determination of the exciton binding energy in conjugated polymers via STM injection measurements. *Phys. Rev. Lett.* (in the press).
30. Strickler, S. J. & Berg, R. A. Relationship between absorption intensity and fluorescence lifetime of molecules. *J. Chem. Phys.* **37**, 814–822 (1962).

Acknowledgements. We thank T. Richardson for the 4TBPPZn samples. D.G.L. thanks Lloyds of London for a Tercentenary research fellowship and D.D.C.B. thanks the Leverhulme Trust for a Leverhulme fellowship. This work was supported by the UK Engineering and Physical Sciences Research Council.

Correspondence and requests for materials should be addressed to D.G.L. (e-mail: d.g.lidzey@sheffield.ac.uk).

A transient liquid-like phase in the displacement cascades of zircon, hafnion and thorite

A. Meldrum*, S. J. Zinkle†, L. A. Boatner* & R. C. Ewing‡

* Oak Ridge National Laboratory, Solid State Division, Oak Ridge, Tennessee 37831-6057, USA

† Oak Ridge National Laboratory, Metals and Ceramics Division, Oak Ridge, Tennessee 37831-6376, USA

‡ The University of Michigan, Department of Nuclear Engineering and Radiological Sciences and the Department of Geological Sciences, Ann Arbor, Michigan 48109-2104, USA

The study of radiation effects in solids is important for the development of 'radiation-resistant' materials for fission-reactor applications¹. The effects of heavy-ion irradiation in the isostructural orthosilicates zircon (ZrSiO_4), hafnion (HfSiO_4) and thorite (ThSiO_4) are particularly important because these minerals are under active investigation for use as a waste form for plutonium-239 resulting from the dismantling of nuclear weapons²⁻⁴. During ion irradiation, localized 'cascades' of displaced atoms can form as a result of ballistic collisions in the target material, and the temperature inside these regions may for a short time exceed the bulk melting temperature. Whether these cascades do indeed generate a localized liquid state⁵⁻⁸ has, however, remained unclear. Here we investigate the irradiation-induced decomposition of zircon and hafnion, and find evidence for formation of a liquid-like state in the displacement cascades. Our results explain the frequent occurrence of ZrO_2 in natural amorphous zircon⁹⁻¹². Moreover, we conclude that zircon-based nuclear waste forms should be maintained within strict temperature limits, to avoid potentially detrimental irradiation-induced amorphization or phase decomposition of the zircon.

Zircon is a commonly occurring accessory mineral found in a wide variety of rock types that typically contains 5–4,000 p.p.m. uranium and up to 2,000 p.p.m. thorium. It has been synthesized with up to 10 wt% plutonium (ref. 13), and pure PuSiO_4 has the tetragonal zircon structure¹⁴. Most natural zircons contain less than 3 mol% hafnium, and the average Zr:Hf ratio in zircon is 40:1—a value which corresponds to the natural abundance of these elements in the Earth's crust. However, hafnion grains with up to 97 mol% Hf have been reported¹⁵ (with the remaining 3% being Zr), and zircon forms a complete solid solution with hafnion. The actinide-bearing

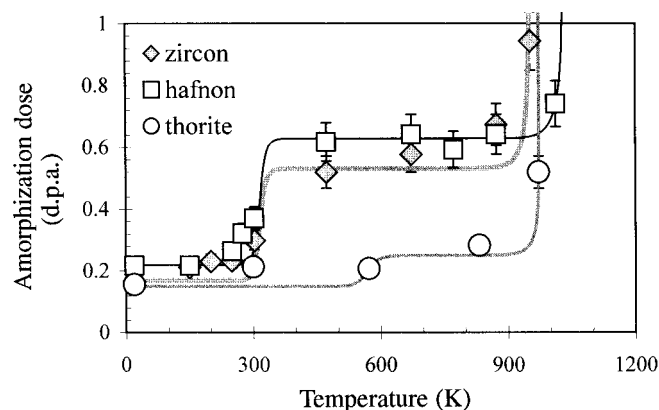


Figure 1 Critical amorphization dose as a function of temperature for zircon, hafnion and thorite. The error bars are 10%, and the lines were obtained by solving an amorphization/recrystallization model which we have developed²⁰. Ion irradiations were performed at the HVEM-Tandem Facility at Argonne National Laboratory using a flux of 1.7×10^{12} ions $\text{cm}^{-2} \text{s}^{-1}$.

mineral thorite (ThSiO_4) is isostructural with zircon and, like zircon, is frequently found to be metamict (amorphous). The lack of intermediate compositions between zircon and thorite suggests a relatively wide miscibility gap¹⁶.

Pure, synthetic zircon, hafnion and thorite crystals were irradiated with 800-keV Kr^+ ions *in situ* in a transmission electron microscope. All of the materials were amorphized at temperatures up to 900 K. The fluence was converted to displacements per atom (d.p.a.) through TRIM-96¹⁷ calculations using atomic displacement energy values of 79 eV (Zr), 23 eV (Si) and 47 eV (O) recently estimated for zircon¹⁸. At low doses, disordered regions ~ 3 nm in diameter, corresponding to individual damage cascades, were observed by high-resolution microscopy. At higher doses, the materials gradually became amorphous, as determined by electron-diffraction. The amorphization dose (D_c) was observed to increase in two stages as a function of temperature (Fig. 1), consistent with previous experiments on zircon¹³ and α -alumina¹⁹.

The temperature dependence of the amorphization dose can be modelled by assuming that the effect of temperature is to cause continuous recrystallization at the cascade boundaries²⁰. For the orthosilicates, the activation energies for irradiation-enhanced recrystallization are 3–3.6 eV. These high values explain why natural specimens of orthosilicates are frequently found to be metamict. In comparison, we have found that the activation energies for the time at invariably crystalline radionuclide-bearing minerals monazite (CePO_4) and apatite ($\text{Ca}_5(\text{PO}_4)_3\text{F}$) are lower by a factor of nearly 3. Using these activation energies for the orthosilicates, a minimum temperature of ~ 700 K is required to ensure the long-term crystallinity of a Pu-loaded zircon waste form.

On increasing the temperature to 950 K (zircon) and $\sim 1,000$ K (hafnion), a new effect was observed. Above these temperatures, the material decomposed under irradiation into the component oxides: tetragonal ZrO_2 or HfO_2 + amorphous SiO_2 . The development of these new phases was determined by electron diffraction (Fig. 2). At 300 K, zircon was amorphized at 0.3 d.p.a. but at 1,050 K zircon and hafnion decomposed directly into their component oxides. In contrast, ThO_2 did not precipitate in the displacement cascades of crystalline thorite at these temperatures. High-resolution electron microscopy confirmed the presence of nanocrystalline ZrO_2 and HfO_2 in the zircon and hafnion specimens (Fig. 3). The observed crystallite size is comparable to the cascade dimensions.

Specimens amorphized at low temperatures were also heated above the temperature at which the radiation-induced phase

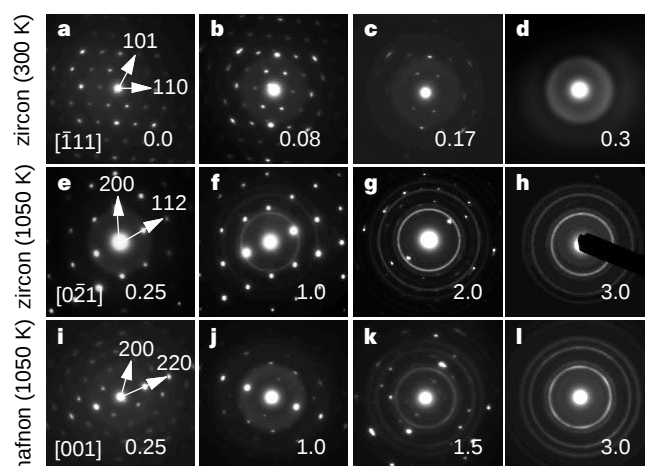


Figure 2 Sequences of electron-diffraction patterns documenting the effects of ion irradiation at different temperatures. **a–d**, Zircon at 300 K; **e–h**, zircon at 1,050 K; and **i–l**, hafnion at 1,050 K. At high temperatures, the gradual development of diffraction 'rings' marked the onset of the phase decomposition. The dose in d.p.a. is given at the bottom right of each diffraction pattern. The arrows point to higher-order spots.

decomposition was observed and held for 20 minutes to determine if the formation of the component oxides was thermally induced. Phase decomposition did not occur in any of the specimens amorphized at low temperature and then held at 1,100 K for 20 minutes. However, when the ion beam was then allowed to impinge on the amorphous specimens, rapid formation of crystalline ZrO_2 and HfO_2 was observed. The component oxide phases did not redissolve when held for 60 minutes at 1,100 K. Specimen heating caused by the ion beam was quantified by irradiating specimens of the ferroelectric phase-change material KNbO_3 under identical conditions and was found to be less than 30 K.

These experiments show that the phase decomposition of zircon and hafnion into their component oxides is nucleated through the effects of the ion irradiation. This has important ramifications in situations where these materials may be exposed to irradiation, particularly at elevated temperatures. For example, the present results imply that a Pu-loaded zircon waste form should be maintained at temperatures well below the phase decomposition temperature in order to avoid the redistribution of Pu and its daughter products between the component oxides. On the basis of these results and on the model derived to plot the lines in Fig. 1, there is a temperature 'window' between approximately 700 and 900 K in which a Pu-doped zircon waste form will neither amorphize nor undergo phase decomposition.

A convincing explanation for the presence of crystalline precipitates at high irradiation temperatures in the displacement cascades of zircon and hafnion and their absence in thorite can be based on the 'thermal spike' model. Recent molecular dynamics simulations suggest that collision cascades have two main stages: an initial ballistic stage during which many atoms are dislodged from their lattice sites, and a subsequent thermal spike phase in which the cascade region attains thermal equilibrium with its surroundings^{6,21,22}. During the thermal spike, the atoms inside the displacement cascade may achieve a liquid-like state (cascade melting), a phenomenon that is of general interest for both fundamental condensed-matter physics and applied technologies such as the ion-beam processing of semiconductors. However, there has been little experimental evidence to substantiate these predictions.

From an examination of the published ZrO_2 - SiO_2 phase diagram²³, crystalline ZrO_2 coexists with an SiO_2 -rich liquid phase at temperatures between 1,960 and 2,670 K for a composition of ZrSiO_4 . Relatively rapid liquid-phase diffusion can occur in this temperature range. A solid solution of crystalline SiO_2 and ZrO_2

occurs below the eutectic solidus temperature of 1,960 K (with correspondingly slower solid-state diffusion parameters), and the solid-state reaction $\text{ZrO}_2 + \text{SiO}_2 \rightarrow \text{ZrSiO}_4$ occurs at 1,950 K. On the basis of the estimated average kinetic energy for the $\sim 1,000$ atoms contained within the individual displacement cascade regions, the effective temperature should exceed the melting point for zircon. According to the thermal spike model, the relatively high quench rate through the two-phase region (crystalline ZrO_2 and SiO_2 -rich liquid), which occurs for low ambient temperatures, would not allow sufficient time for nucleation of the crystalline ZrO_2 phase. However, decreasing the cooling rate at temperatures in the two-phase region between 2,670 and 1,960 K by increasing the ambient temperature could allow the nucleation of tetragonal ZrO_2 in the displacement cascade region.

The HfO_2 - SiO_2 phase diagram²³ shows a two-phase region (crystalline HfO_2 plus SiO_2 -rich liquid) from 2,870 to 2,020 K for a composition of HfSiO_4 . A peritectic reaction to form hafnion occurs at 2,020 K. The peritectic solidus temperature in hafnion is 63 K higher than the eutectic solidus temperature in zircon; this temperature difference is close to the difference in minimum irradiation temperatures for the formation of crystalline HfO_2 versus ZrO_2 crystallites in the displacement cascades of the two materials (1,000 K versus 950 K, respectively).

The published ThO_2 - SiO_2 phase diagram²³ shows a two-phase region (crystalline ThO_2 plus SiO_2 -rich liquid) from $\sim 2,470$ to 2,250 K for a composition of ThSiO_4 . Thorite is formed by a peritectic reaction with a solidus temperature of 2,250 K. The relatively small temperature range between the liquidus and the solidus, and also the relatively high solidus temperature for thorite compared to hafnion and zircon, are not favourable for the precipitation of ThO_2 during cooling of a displacement cascade. Therefore, precipitation of ThO_2 in the displacement cascade requires higher ambient temperatures, and the transformation to the component oxides occurs less readily than for the cases of zircon and hafnion.

The irradiation-induced nucleation of new phases in crystalline or amorphous zircon and hafnion is consistent with the concept of thermal spikes. Within the thermal spike model, the temperature in the cascade region is characterized by a cooling rate $\exp(-t/\tau)$, where t is time and the cooling time constant τ for metals such as nickel with strong electron-phonon coupling is ~ 1 ps (refs 24, 25). As the temperature profile is inversely proportional to the thermal conductivity, and the thermal conductivity of the orthosilicates at $T > 1,800$ K is approximately an order of magnitude lower than that of nickel, the cooling time constant for the orthosilicates should be at least 10 ps. The magnitude of the diffusion coefficient necessary to produce solute segregation within the displacement cascade region can be estimated to be $D \approx x^2/t \approx 1 \times 10^{-8} \text{ m}^2 \text{ s}^{-1}$ where $x = 1.5$ nm is the observed crystallite radius (Fig. 3), the cooling time constant τ is 10 ps, and $t \approx 3\tau$ is the thermal spike lifetime. This estimate is similar to typical liquid diffusion coefficients²⁶.

Crystalline ZrO_2 has been observed to precipitate in metamict natural zircons^{9,10,12} and in ion-beam-amorphized zircons heated to $\sim 1,200$ K (ref. 27). Zirconia-rich and silica-rich microdomains were proposed to exist in metamict zircon, evidence for which has recently been obtained by ion microprobe techniques¹¹. The present results imply that these domains may form during the cooling of the α -recoil cascades in the U and Th decay series. ZrO_2 - and SiO_2 -rich 'nanodomains' should occur as fine-scale fluctuations in the amorphous material, and these nanodomains could represent nuclei for the subsequent precipitation of ZrO_2 in metamict natural zircons. These fluctuations may lead to a fine-scale redistribution of U and Pb contained in natural zircon used for geochronology^{28,29}.

In contrast, irradiation with light ions (for example, He^+) does not produce displacement cascades, but instead results in the formation of isolated point defects. So the phase decomposition observed in the present experiments is not expected to occur during irradiation with light ions or as a result of subsequent thermal

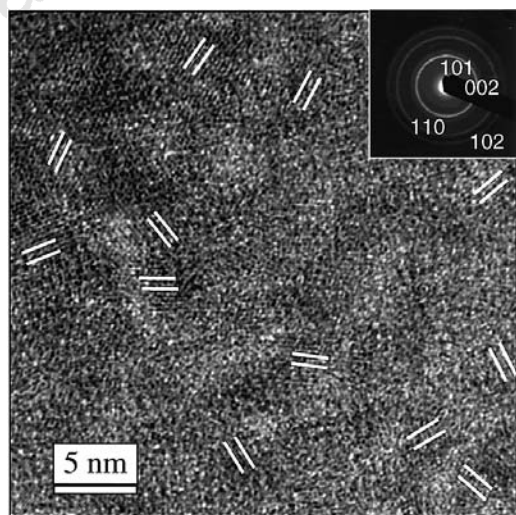


Figure 3 High-resolution TEM image of zircon irradiated at 1,050 K (dose, 3 d.p.a.), showing randomly orientated ZrO_2 nanocrystals (highlighted). The corresponding electron-diffraction pattern is given in the inset. Energy dispersive X-ray spectroscopy was used to confirm the presence of an amorphous SiO_2 component.

annealing. The use of light ions is, therefore, suggested for the formation of optical waveguides³⁰ in zircon. □

Received 13 February; accepted 8 June 1998.

- Seitz, F. Radiation effects in solids. *Phys. Today* **5**, 6–9 (1952).
- Burakov, B. E. *et al.* in *Scientific Basis for Nuclear Waste Management XIX* (eds Murphy, W. M. & Knecht, D. A.) 33–40 (Plenum, New York, 1996).
- Ewing, R. C. *et al.* Zircon: A host phase for the disposal of weapons plutonium. *J. Mater. Res.* **10**, 243–246 (1995).
- Weber, W. J. *et al.* in *Scientific Basis for Nuclear Waste Management XIX* (eds Murphy, W. M. & Knecht, D. A.) 25–32 (Plenum, New York, 1996).
- Brinkman, J. A. On the nature of radiation damage in metals. *J. Appl. Phys.* **25**, 961–970 (1954).
- Diaz de la Rubia, T. *et al.* Molecular dynamics simulation of displacement cascades in Cu and Ni: Thermal spike behavior. *J. Mater. Res.* **4**, 579–586 (1989).
- Wang, S. X. *et al.* Ion beam-induced amorphization in the MgO–Al₂O₃–SiO₂ system: Part II—An empirical model. *J. Non-Cryst. Solids* (in the press).
- Sales, B. C. *et al.* Structural differences between the glass state and ion-beam-amorphized states of lead pyrophosphate. *J. Non-Cryst. Solids* **126**, 179–193 (1990).
- Ellsworth, S. *et al.* Energetics of radiation damage in natural zircon (ZrSiO₄). *Phys. Chem. Mineral.* **21**, 140–149 (1994).
- Farges, F. The structure of metamict zircon: A temperature-dependent EXAFS study. *Phys. Chem. Mineral.* **20**, 504–514 (1994).
- McLaren, A. C. *et al.* The microstructure of zircon and its influence on the age determination from Pb/U isotopic ratios measured by ion microprobe. *Geochim. Cosmochim. Acta* **58**, 993–1005 (1994).
- Vance, E. R. & Anderson, B. W. *Mineral Mag.* **38**, 605–613 (1972).
- Weber, W. J. *et al.* The radiation-induced crystalline-to-amorphous transition in zircon. *J. Mater. Res.* **9**, 688–698 (1994).
- Keller, C. Untersuchungen über die germanate und silikate des typs ABO₄ der vierwertigen elemente thorium bis americium. *Nukleonik* **5**, 41–48 (1963).
- Correia Neves, J. M. *et al.* High-hafnium members of the zircon-hafnium series from the granite pegmatites of Zambezia, Mozambique. *Contrib. Mineral. Petrol.* **48**, 73–80 (1974).
- Speer, J. A. in *Reviews in Mineralogy* Vol. 5, *Orthosilicates* (ed. Ribbe, P. H.) 67–132 (Mineralogical Soc. Am. Washington DC, 1982).
- Ziegler, J. F. *Transport and Range of Ions in Matter* Version 96.01 (IBM-Research, Yorktown, NY, 1996).
- Weber, W. J. *et al.* Radiation effects in crystalline ceramics for the immobilization of high-level nuclear waste and plutonium. *J. Mater. Res.* (in the press).
- Abe, H. *et al.* Amorphization in aluminum oxide induced by ion irradiation. *Nucl. Instrum. Meth. Phys. Res.* **B127/128**, 170–175 (1997).
- Meldrum, A. *et al.* Heavy-ion irradiation effects in the ABO₄ orthosilicates: decomposition, amorphization, and recrystallization. *Phys. Rev. B* (submitted).
- Heinisch, H. L. & Singh, B. N. On the structure of irradiation induced collision cascades in metals as a function of recoil energy and crystal structure. *Phil. Mag.* **67**, 407–424 (1993).
- Bacon, D. J. & Diaz de la Rubia, T. Molecular dynamics computer simulations of displacement cascades in metals. *J. Nucl. Mater.* **216**, 275–290 (1994).
- Levin, E. M. & McMurdie, H. F. *Phase Diagrams for Ceramists*, 1975 Supplement (Am. Ceramic Soc., Columbus, OH, 1975).
- Caro, M. *et al.* Thermal behavior of radiation damage cascades via the binary collision approximation: Comparison with molecular dynamics results. *J. Mater. Res.* **5**, 2652–2657 (1990).
- Koponen, I. Energy transfer between electrons and ions in dense displacement cascades. *Phys. Rev. B* **47**, 14011–14019 (1993).
- Kirkaldy, J. S. & Young, D. J. *Diffusion in the Condensed State* (Institute of Metals, London, 1987).
- Weber, W. J. Radiation-induced defects and amorphization in zircon. *J. Mater. Res.* **5**, 2687–2697 (1990).
- Buick, R. *et al.* Record of emergent continental crust ~3.5 billion years ago in the Pilbara craton of Australia. *Nature* **375**, 574–577 (1995).
- Froude, D. O. *et al.* Ion microprobe identification of 4,100–4,200 Myr-old terrestrial zircons. *Nature* **304**, 616–618 (1983).
- Babsail, L. *et al.* Helium-ion implanted waveguides in zircon. *Nucl. Instrum. Meth. Phys. Res.* **B59/60**, 1219–1222 (1991).

Acknowledgements. This work was sponsored by the Division of Materials Sciences, Basic Energy Sciences, and the Environmental Management Sciences Program, US-DOE, with Lockheed Martin Energy Research Corporation.

Correspondence and requests for materials should be addressed to A.M. (e-mail: al-m@worldnet.att.net).

Contribution of hurricanes to local and global estimates of air–sea exchange of CO₂

Nicholas R. Bates*, Anthony H. Knap* & Anthony F. Michaels†

* Bermuda Biological Station For Research, Inc., Ferry Reach, GEO1, Bermuda
† Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089-0371, USA

The effect of hurricanes on the thermal and physical structure of the upper ocean has been described¹ but their influence on the ocean carbon cycle and the exchange of carbon between ocean and atmosphere is not well understood. Here we present observations from the Sargasso Sea, before and after hurricane Felix in summer 1995, that show a short-lived (2–3 weeks) surface seawater cooling of about 4 °C, and a decrease in seawater partial pressure of CO₂

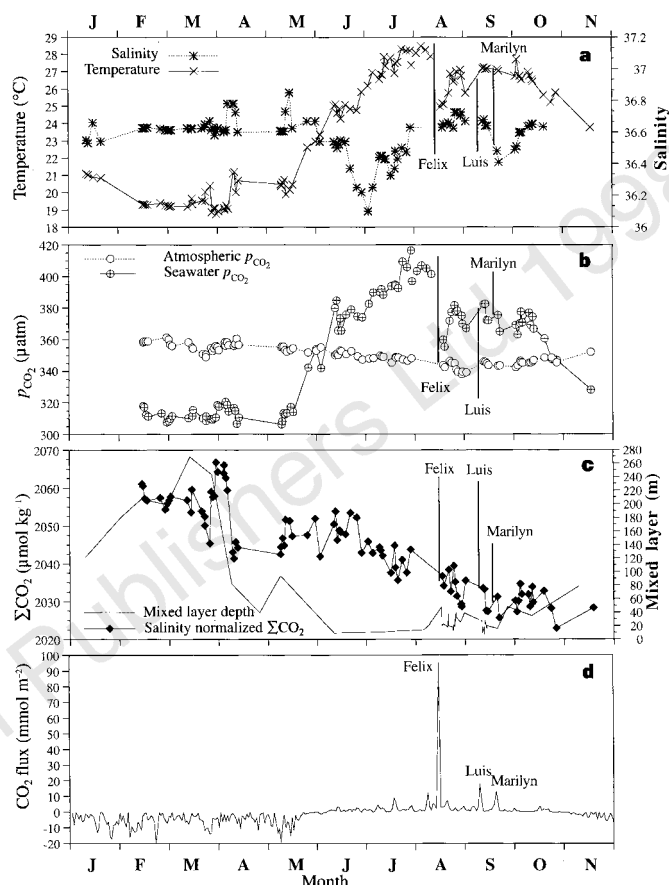


Figure 1 Time-series of surface properties in the Sargasso Sea near Bermuda. Each datum point represents the mean daily value of each property. **a**, Sea surface temperature and salinity. Thermosalinograph readings of these properties were measured every 10 s from sea water supplied by the ship's bow intake (~1.5 m deep). The last five data points before hurricane Felix represent 3-day averages of sea surface temperature determined by high-resolution AVHRR infrared data. **b**, Surface seawater and atmospheric p_{CO_2} . Surface seawater p_{CO_2} values for the last five data points before hurricane Felix represent 3-day averages of p_{CO_2} calculated from p_{CO_2} –temperature relationships observed from June and July. Atmospheric and surface water p_{CO_2} was measured using an automated infrared system (designed by T. Takahashi & D. W. Chipman, Lamont Doherty Earth Observatory), similar to the gas-chromatographic p_{CO_2} system described in ref. 25. The system alternately analyses dried atmospheric air and air equilibrated with sea water from a 40-l shower-head equilibrator. The measurement was calibrated every 30 minutes with 4 standards of known CO₂ content, which have been referenced to World Meteorological Organization (WMO) CO₂ standards. The analysis has a precision of better than $\pm 1 \mu\text{atm}$. Surface water p_{CO_2} data was collected at intervals of ~4 minutes, while atmospheric p_{CO_2} was determined every 25 min. Equilibrator temperatures and atmospheric pressure were measured every 2 min. Differences between *in situ* surface temperature and equilibrator temperatures were $<0.06^\circ\text{C}$, requiring no correction to equilibrator p_{CO_2} data. **c**, Total CO₂ (ΣCO_2) normalized to a constant salinity of 36.6 ($\mu\text{mol kg}^{-1}$) and mixed layer depth (m). ΣCO_2 was calculated from seawater p_{CO_2} and alkalinity using appropriate thermodynamic considerations^{26,27} and dissociation constants^{28,29}. ΣCO_2 is the total concentration of $[\text{HCO}_3^-]$, $[\text{CO}_3^{2-}]$, $[\text{CO}_2]$ and $[\text{H}_2\text{CO}_3]$ in seawater. Alkalinity was calculated from salinity as alkalinity was well correlated with salinity in the Sargasso Sea³⁰. The precision of calculated ΣCO_2 and alkalinity was $\pm 3 \mu\text{mol kg}^{-1}$ (N.R.B., unpublished data). **d**, Daily fluxes of CO₂ in the Sargasso Sea. Fluxes were calculated using daily averaged winds and the piston velocity–wind speed relationships of ref. 18. Daily averaged wind speed data was collected in the vicinity of the sampling region at the Bermuda Testbed Mooring¹⁵ and at the US Naval Air Station on the island of Bermuda (8 m height).

Table 1 Comparison of CO₂ fluxes due to hurricane and seasonal forcing in the Sargasso Sea

Date	Hurricane	Day of year	Δp_{CO_2} (μatm)	Mean winds (m s^{-1})	CO ₂ flux; ref. 17* (mmol CO ₂ m ⁻²)	CO ₂ flux; ref. 18† (mmol CO ₂ m ⁻²)	CO ₂ flux; ref. 17‡ (mmol CO ₂ m ⁻²)
14 August 1995	Felix	225	+65	14.9	+15	+32	+40
15 August 1995	Felix	226	+65 to +15	22.9	+25	+76	+95
					Total +40	+108	+135
2 Sept. 1995	Luis	251	+34	14.3	+7	+14	+18
3 Sept. 1995	Luis	252	+34	9.9	+4	+7	+9
10 Sept. 1995	Marilyn	261	+31	13.5	+6	+10	+10
11 Sept. 1995	Marilyn	261	+31	9.1	+4	+8	+7
Total hurricane efflux (% of total summer efflux)			NA	NA	+61 (40%)	+145 (44%)	+182 (44%)
Total summer efflux (1995)		156–301	NA	NA	+151	+331	+417
Total summer efflux (1994)		158–297	NA	NA	+78	+163	+205

See Methods for details; positive numbers in table denote flux into atmosphere.

* Equations (2)–(4) in Methods.

† Equation (5) in Methods.

‡ Equation (6) in Methods.

by about 60 μatm . Despite the localized decrease in seawater partial pressure of CO₂, strong winds during the passage of hurricane Felix increased the efflux of CO₂ from ocean to atmosphere. We estimate that hurricane Felix and two other hurricanes increased the summertime efflux of CO₂ into the atmosphere over this part of the Sargasso Sea by nearly 55%. We estimate that hurricanes contribute to the global ocean-to-atmosphere flux of CO₂ by between +0.04 to +0.51 PgC (10¹⁵ gC) per year. Such hurricane-forced effluxes are quantitatively significant compared to regional (14° to 50° N zone)² and global effluxes^{2,3}. Hurricanes therefore exert an important influence on ocean–atmosphere CO₂ exchange and the inferred⁴ year-to-year variability of CO₂ fluxes over the subtropical oceans.

Surface seawater partial pressure of CO₂ (p_{CO_2}) is controlled by several factors, such as temperature, salinity, changes in total carbon dioxide (ΣCO_2) or alkalinity, ocean mixing, gas exchange and primary production. For example, cooling of sea water reduces p_{CO_2} by ~4.1–4.25% per °C (refs 5, 6), marine phytoplankton blooms reduce both p_{CO_2} and ΣCO_2 (refs 7, 8), while mixing upwards of CO₂ from CO₂-rich subsurface water tends to increase surface p_{CO_2} (refs 9, 10). However, the influence of hurricanes (tropical cyclones) on seawater p_{CO_2} and the air–sea gas exchange of CO₂ is unknown.

In 1995, we observed the effects of hurricanes on sea surface temperature, salinity, and seawater and atmospheric p_{CO_2} in the Sargasso Sea. Our detailed time series of measurements were collected during numerous (>55), 1–5 day cruises in the vicinity of hydrostation S¹¹ (32° 10' N, 64° 30' W) and the Bermuda Atlantic Time-series Study (BATS) site^{12,13} (31° 50' N, 64° 10' W), ~25–45 km southwest of Bermuda. Located in the western North Atlantic Ocean, this region has a well documented seasonal cycle of mixing, temperature and biological production^{11–13}. In wintertime, deep convective mixing (~120 to >250 m) brings nutrients into the euphotic zone, stimulating a short-lived spring phytoplankton bloom. From April to May, the mixed-layer shoals to ~5–20 m deep and a warm surface layer is present until deep mixing begins again in October¹³. Seasonal variability of ΣCO_2 , seawater and atmospheric p_{CO_2} is 40–50 $\mu\text{mol kg}^{-1}$, 90–110 μatm and 15–25 μatm , respectively^{9,10}.

On 14 and 15 August 1995, hurricane Felix passed 50 km south of Bermuda, directly over the sampling region. This hurricane was quite large (~350 km in diameter) and had sustained winds of ~40–45 m s^{-1} . For several weeks before the storm, surface temperatures remained close to 28 °C (Fig. 1a). In the wake of hurricane Felix, a surface of cooling of ~4 °C was observed, the trace of which was visible by satellite imagery for several weeks¹⁴. The surface cooling resulted from upward mixing of subsurface waters during deepening of the mixed layer¹⁵, similar to thermal effects of hurricanes previously observed in the North Atlantic¹. A large

change in seawater p_{CO_2} was also observed. For several weeks before the storm, surface p_{CO_2} ranged from 400 to 420 μatm (Fig. 1b). Three days after passage of hurricane Felix, we observed that seawater p_{CO_2} was ~60 μatm lower, a large decrease considering that the annual changes were typically 80–100 μatm (refs 9, 10, 16). The change in p_{CO_2} resulted primarily from temperature changes. We estimated, according to empirical and theoretical p_{CO_2} temperature relationships, that a cooling of 4 °C should result in a decrease of p_{CO_2} by 50–55 μatm . Other potential causes such as upwelling or salinity changes were ruled out because no clear differences between pre- and post-storm surface ΣCO_2 (Fig. 1c), vertical gradients of ΣCO_2 , salinity (Fig. 1a) or alkalinity were observed. After passage of hurricane Felix, seawater p_{CO_2} levels rebounded by 20 μatm several weeks after the storm, but never returned to pre-storm levels. In September, hurricanes Luis and Marilyn passed to the west of Bermuda at distances of 200–250 km from the sampling area. Due to their distance from Bermuda, we observed no changes in surface temperatures, p_{CO_2} , ΣCO_2 or salinity.

The direction and rate of air–sea gas exchange of CO₂ is a function of piston velocity, which is related to wind speed, and the concentration difference between seawater and atmospheric p_{CO_2} (Δp_{CO_2} ; refs 17, 18)) (see Table 1 for details). During typical summertime conditions in the Sargasso Sea, wind speeds are low (<5 m s^{-1}) and seawater p_{CO_2} levels are higher than atmospheric p_{CO_2} levels (that is, positive Δp_{CO_2} values) causing a small flux of CO₂ from ocean to atmosphere^{9,10}. Before and after hurricane Felix, Δp_{CO_2} values were ~+65 μatm and ~+15 μatm respectively, conditions favourable for continued flux of CO₂ from ocean to atmosphere.

We calculated the flux of CO₂ during hurricane Felix using hourly averaged winds and several piston velocity–wind speed relationships^{17,18}. Although we only have pre- and post-storm seawater and atmospheric p_{CO_2} values, we calculated hourly fluxes over the two days of the storm (Table 1). We assumed that seawater p_{CO_2} remained at 410 μatm until the height of the storm, and then declined linearly to 360 μatm at the end of the storm. To calculate CO₂ fluxes, we assumed that Δp_{CO_2} values were +65 μatm at the beginning of the storm, +15 μatm at the end and that atmospheric p_{CO_2} decreased by ~20 μatm in response to a transient 6% drop in atmospheric pressure (from ~1,020 to 965 mbar). Several piston velocity–wind speed relationships were used in the calculation to bound the range of uncertainty of air–sea gas exchange. Although piston velocities have not been measured above wind speeds of ~18 m s^{-1} , we calculated fluxes for stronger winds assuming that the linear relationships of gas exchange at wind speeds between 10 and 18 m s^{-1} hold for higher wind speeds. We estimated that the loss of CO₂ from the ocean due to hurricane Felix ranged between 40 and 135 mmol CO₂ m⁻² (Fig. 1d) (Table 1; details of the methods used to estimate air–sea fluxes are given in Methods), compared with the

Table 2 The influence of hurricanes on CO₂ fluxes, 1983 to 1992

Year	Total time of all storms (d)		Diameter of storm (km)	Flux of CO ₂ (Pg C)	
	Hurricane	Tropical storm	Hurricane or tropical storm	Δp _{CO₂} (+50 μatm)	Δp _{CO₂} (+100 μatm)
All oceanic basins 40° N to 40° S)					
1992	248.7	525.3	100	+0.067	+0.135
			150	+0.109	+0.218
			250	+0.255	+0.509
1991	172.0	411.5	100	+0.050	+0.100
			150	+0.078	+0.157
			250	+0.185	+0.370
1990	208.1	473.4	100	+0.050	+0.117
			150	+0.093	+0.187
			250	+0.219	+0.439
1989	201.4	449.8	100	+0.043	+0.086
			150	+0.055	+0.110
			250	+0.146	+0.292
1988	149.6	348.5	100	+0.042	+0.085
			150	+0.068	+0.136
			250	+0.159	+0.318
1987	133.8	404.8	100	+0.044	+0.089
			150	+0.067	+0.134
			250	+0.159	+0.319
1986	174.8	457.5	100	+0.053	+0.107
			150	+0.083	+0.166
			250	+0.195	+0.390
1985	170.5	481.3	100	+0.054	+0.109
			150	+0.083	+0.166
			250	+0.197	+0.394
1984	172.0	481.8	100	+0.055	+0.109
			150	+0.084	+0.167
			250	+0.198	+0.396
1983	157.7	401.3	100	+0.047	+0.095
			150	+0.074	+0.148
			250	+0.174	+0.348
Northern Hemisphere only (0° to 40° N)					
1992	188.8	419.3	100	+0.052	+0.105
			150	+0.074	+0.148
			250	+0.174	+0.348
1987	90.3	227.8	100	+0.027	+0.054
			150	+0.042	+0.084
			250	+0.093	+0.199
Range of fluxes due to hurricanes (from values above)					
40° N–40° S (latitudinal band of hurricanes)			This study	+0.042 to 0.509	
Comparison to global ocean fluxes of CO ₂					
Global ocean			Ref. 18	–0.290 to –0.690	
Global ocean			Ref. 22	–1.170 to –1.340	
Comparison to regional ocean fluxes of CO ₂					
14° N–50° N (northern temperate zone)			Ref. 22	–0.680 to –0.720	
14° N–14° S (tropical zone)			Ref. 22	+1.040 to +1.040	
14° S–50° S (southern temperate zone)			Ref. 22	–0.870 to –0.920	

See Methods for details; positive numbers in table represent flux into atmosphere.

total loss of CO₂ from the ocean of +151 to +417 mmol CO₂ m^{–2} during the entire summer. If we include the effect of strong winds associated with hurricanes Luis and Marilyn, we calculate that the contribution of hurricanes to the flux of CO₂ in the region was ~61–182 mmol CO₂ m^{–2}, nearly 45% of the total efflux during the summer (Table 1). Although the effect of the three hurricanes lasted for a relatively short period, the net local effect was to double the summertime efflux of CO₂. The contribution of hurricanes to the summertime efflux of CO₂ in 1995 was greater than the total summertime flux in 1994 (Table 1).

Hurricanes clearly have a considerable influence on local rates of air–sea gas exchange of CO₂, but how important are they to the global carbon cycle? To answer this question, we estimated the net yearly contribution of hurricanes using data from the Global Tropical Extra-tropical Cyclone Climatic Atlas (GTECCA), issued by the National Climatic Data Center, which provides information about the intensity track and duration of all storms. We calculated the total number of days that hurricane or tropical storms were present over the ocean surface each year from 1983 to 1992 (Table 2; Details of the methods used to estimate air–sea fluxes due to hurricanes and tropical storms are given in Methods). Although data about storm size, wind field patterns and Δp_{CO_2} conditions

were lacking, we estimated ocean-to-atmosphere CO₂ flux for hurricanes and tropical storms by making reasonable assumptions about average storm area, wind speed and Δp_{CO_2} values (see Methods for details). To account for natural variability in the size and intensity of storms, we used three different average hurricane and tropical storm (with mean winds of 45 and 25 m s^{–1}, respectively) diameters of 100, 150 and 250 km to calculate the oceanic area influenced by these storms. Tropical storm-force winds (25 m s^{–1} around hurricanes and 15 m s^{–1} around tropical storms) were assumed to extend beyond the centre of the storm by twice these storm diameters. Wind speed-transfer velocity relationships^{17,18} were assumed to hold at higher wind speeds. As surface values are typically supersaturated in subtropical regions^{19–21}, we used two Δp_{CO_2} values of +50 and +100 μatm for the analysis.

Our simple calculations suggest that hurricanes and tropical storms in the latitudinal band 40° S to 40° N should contribute to the ocean-to-atmospheric flux of CO₂ by between +0.042 and +0.509 Pg C yr^{–1} (Table 2). If we compare the yearly flux estimates for storms with a Δp_{CO_2} of +100 μatm and diameter of 300 km (see Table 2), there was significant year-to-year variability of fluxes (for example, a minimum of +0.292 Pg C yr^{–1} in 1989 and maximum of +0.509 Pg C yr^{–1} in 1992). The CO₂ flux due to hurricanes (+0.042

to $+0.509 \text{ Pg C yr}^{-1}$) was significant when compared to estimates^{2,3} of the global atmosphere-to-ocean CO_2 flux (that is, -0.24 to $-1.34 \text{ Pg C yr}^{-1}$; Table 2). Within the Northern Hemisphere (0° to 40°N), we estimated that hurricanes contributed to an ocean-to-atmosphere CO_2 flux in the range of $+0.02$ to $+0.20 \text{ Pg C}$ (compare 1992 and 1987; Table 2). This range of flux is comparable to estimates^{2,3} of the regional ocean-to-atmosphere CO_2 fluxes ($+0.36$ to $+0.72 \text{ Pg C yr}^{-1}$; Table 2) in the northern temperate zone (14°N to 50°N). Thus it seems that year-to-year differences in the number and intensities of hurricanes are an important variable for controlling the magnitude of CO_2 fluxes within the Northern Hemisphere temperate zone. Although we expect that changes in surface-to-deep mixing remains the primary control of ocean uptake of CO_2 over multi-year, decadal timescales²², CO_2 fluxes due to hurricanes provide an additional secondary feedback mechanism that is not accounted for in present global carbon cycle and climate models. Recent studies suggest that although there are still many uncertainties about predicting whether climate change will alter the frequency, intensity and geographical distribution of tropical cyclones, there may be a modest 10–20% increase in tropical cyclone intensity²³, which in turn would increase the importance of the ocean-to-atmosphere CO_2 flux of hurricanes. □

Methods

Calculation of daily, seasonal and annual air-sea fluxes of CO_2 (Table 1).

Three formulations of the piston velocity^{17,18} were used to bound the range of uncertainties associated with the definition of the CO_2 gas transfer coefficient (that is, minimum and maximum estimates of the CO_2 flux). The net flux of CO_2 (F) was calculated from the gas exchange equation:

$$F = ks(\Delta p_{\text{CO}_2}) \quad (1)$$

where k is the transfer velocity, s is the solubility of CO_2 and Δp_{CO_2} is the difference in partial pressure between atmosphere and seawater surface p_{CO_2} . Two relationships were used to provide upper and lower error estimates for the flux calculation. The lower limit is based on the equations in ref. 17, which were derived mainly from wind-tunnel and lake experimental data:

$$k = 0.17u(600/\text{Sc})^{2/3} \quad \text{where } u \leq 3.6 \text{ m s}^{-1} \quad (2)$$

$$k = ((2.8u) - 9.65)(600/\text{Sc})^{1/2} \quad \text{where } 3.6 > u \leq 13 \text{ m s}^{-1} \quad (3)$$

$$k = (5.9u - 49.3)(600/\text{Sc})^{1/2} \quad \text{where } u > 13 \text{ m s}^{-1} \quad (4)$$

where u is wind speed at 10 m above mean sea level and Sc is the Schmidt number for CO_2 . The upper limit is based on two equations from ref. 18, which were based on a quadratic dependency between wind speed and transfer velocity:

$$k = 0.31u^2(\text{Sc}/660)^{-0.5} \quad (5)$$

$$k = 0.38u^2(\text{Sc}/660)^{-0.5} \quad (6)$$

Sc was calculated using the equations from ref. 18, while solubility of CO_2 , s , was calculated from the observed temperature and salinity using the equations from ref. 24. Positive numbers denote flux into the atmosphere.

Calculation of the potential contribution of hurricanes and tropical storms to the flux of CO_2 from ocean to atmosphere (Table 2). These calculations were done for each year between 1983 and 1992. Flux calculations were made for all hurricanes and tropical storms within the latitudinal band 40°S to 40°N and in selected years for the Northern Hemisphere only (0° to 40°N). The range of fluxes due to hurricanes and tropical storms were also compared to global estimates of flux reported in refs 2 and 3, and regional fluxes from ref. 3. The flux of CO_2 is expressed as Pg of C . Methods and assumptions used for the calculations were as follows. First, we calculated the total time that hurricane and tropical storms were present over the global ocean each year. This data was compiled from information provided by the Tropical Prediction Centre about the history of individual storms in all active basins (North Atlantic Ocean, northeast and northwest Pacific Ocean, North Indian Ocean, southwest Pacific Ocean and South Indian Ocean). Second, we assumed an average size, area and wind speed for all storms. To account for the large variability in size and intensity of storm, we calculated surface area for each hurricane or tropical

storm assuming three different diameters (100, 150 and 250 km). Within each hurricane or tropical storm, we used mean winds of 45 and 25 m s^{-1} , respectively. A wind field of 25 and 15 m s^{-1} for hurricane and tropical storm, respectively, was assumed to extend beyond the centre of each storm by double the storm diameter (200, 300 and 500 km). Third, flux of CO_2 was calculated from the wind speed–transfer velocity and solubility equations from ref. 18 assuming a surface seawater temperature of 28°C and salinity of 35. Surface Δp_{CO_2} values are typically supersaturated but highly variable ($\sim +20$ to $\sim +150 \mu\text{atm}$)^{19–21} in ocean regions where hurricanes and tropical storms occur. Therefore, we used two different average Δp_{CO_2} values of $+50$ and $+100 \mu\text{atm}$, respectively, to bound the range of uncertainty about surface conditions. Last, total fluxes of CO_2 per year were calculated using average storm area (three different diameters) and wind speed.

Received 14 July 1997; accepted 26 May 1998.

- Stramma, L., Cornillon, P. & Price, J. F. Large diurnal sea surface temperature variability: satellite and *in situ* measurements. *J. Geophys. Res.* **91**, 5031–5035 (1986).
- Takahashi, T. *et al.* Global air-sea flux of CO_2 : An estimate based on measurements of sea-air $p\text{CO}_2$ difference. *Proc. Natl Acad. Sci. USA* **94**, 8292–8299 (1997).
- Tans, P. P., Fung, I. Y. & Takahashi, T. Observational constraints on the global atmospheric CO_2 budget. *Science* **247**, 1431–1438 (1990).
- Ciais, P. *et al.* Partitioning of ocean and land uptake of CO_2 as inferred by $\delta^{13}\text{C}$ measurements from the NOAA Climate Monitoring and Diagnostics Laboratory Global Air Sampling Network. *J. Geophys. Res.* **100**(D3), 5051–5070 (1995).
- Takahashi, T., Olafsson, J., Goddard, J., Chipman, D. W. & Sutherland, S. C. Seasonal variation of CO_2 and nutrients in the high-latitude surface oceans: a comparative study. *Glob. Biogeochem. Cycles* **7**, 843–878 (1993).
- Millero, F. J. Thermodynamics of the carbonic acid system in seawater. *Geochim. Cosmochim. Acta* **59**, 661–677 (1995).
- Codispoti, L. A., Friederich, G. E. & Hood, D. W. Variability in the inorganic carbon system over the southeastern Bering Sea shelf during spring 1980 and spring–summer 1981. *Continental Shelf Res.* **5**, 133–160 (1986).
- Watson, A. J., Robins, C., Robertson, J. E., LeB. Williams, P. J. & Fasham, M. J. R. Spatial variability in the sink for atmospheric carbon dioxide in the North Atlantic. *Nature* **350**, 50–53 (1991).
- Bates, N. R., Michaels, A. F. & Knap, A. H. Alkalinity changes in the Sargasso Sea: Geochemical evidence of calcification? *Mar. Chem.* **51**, 347–358 (1996).
- Bates, N. R., Michaels, A. F. & Knap, A. H. Seasonal and interannual variability of the oceanic carbon dioxide system at the U.S. JGOFS Bermuda Atlantic Time-series Site. *Deep-Sea Res.* **42**, 347–383 (1996).
- Menzel, D. W. & Ryther, J. H. The annual cycle of primary production in the Sargasso Sea off Bermuda. *Deep-Sea Res.* **6**, 351–367 (1960).
- Lohrenz, S. E. *et al.* Seasonal variability in primary production and particle flux in the northwestern Sargasso Sea: U.S. JGOFS Bermuda Atlantic Time-series Study. *Deep-Sea Res.* **39**, 1373–1391 (1992).
- Michaels, A. F. *et al.* Seasonal patterns of ocean biogeochemistry at the U.S. JGOFS Bermuda Atlantic Time-series Study site. *Deep-Sea Res.* **41**, 1013–1038 (1994).
- Nelson, N. B. The wake of Hurricane Felix. *Int. J. Remote Sensing* **17**, 2893–2895 (1996).
- Dickey, T. D. *et al.* Preliminary results from the Bermuda Testbed Mooring Program. *Deep-Sea Res.* **43** (in the press).
- Bates, N. R., Takahashi, T., Chipman, D. W. & Knap, A. H. Variability of $p\text{CO}_2$ on diel to seasonal timescales in the Sargasso Sea near Bermuda. *J. Geophys. Res.* (in the press).
- Liss, P. S. & Merlivat, L. in *The Role of Air-Sea Exchange in Geochemical Cycling* (ed. Buat-Menard, P.) 113–128 (NATO ASI Ser. C 185, Springer-Verlag, Berlin, 1986).
- Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean. *J. Geophys. Res.* **97**, 7373–7380 (1992).
- Feely, R. A. *et al.* The effect of tropical instability waves on CO_2 distributions along the equator in the eastern equatorial Pacific during the 1992 ENSO event. *Geophys. Res. Lett.* **21**, 277–280 (1994).
- Feely, R. A. *et al.* CO_2 distributions in the equatorial Pacific during the 1991–1992 ENSO event. *Deep-Sea Res.* **42**, 387–409 (1995).
- Wanninkhof, R. *et al.* Seasonal and lateral variations in carbon chemistry of surface water in the eastern equatorial Pacific during 1992. *Deep-Sea Res.* **42**, 387–409 (1995).
- Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**, 245–249 (1998).
- Henderson-Sellers, A. & Zhang, H. *Tropical Cyclones and Global Climate Change* (WMO/CAS/TMRP Committee on Climate Change Assessment (Project TC-2), World Meteorological Organization Rep. New York, NY, 1997).
- Weiss, R. F. Carbon dioxide in water and seawater; the solubility of a non-ideal gas. *Mar. Chem.* **2**, 203–215 (1974).
- Chipman, D. W., Marra, J. & Takahashi, T. Primary production at 47°N and 20°W in the North Atlantic Ocean: a comparison between the ^{14}C incubation method and the mixed layer carbon budget. *Deep-Sea Res.* **40**, 151–169 (1993).
- Peng, T.-H., Takahashi, T., Broecker, W. S. & Olafsson, J. Seasonal variability of carbon dioxide, nutrients and oxygen in the northern North Atlantic surface water: observations and a model. *Tellus B* **39**, 439–458 (1987).
- Dickson, A. G. & Goyet, C. (eds) *Handbook of Methods for the Analysis of the Various Parameters of the Carbon Dioxide System in Seawater* Version 2 (US Department of Energy CO_2 Science Team Report, Washington DC, 1994).
- Goyet, C. & Poisson, A. P. New determination of carbonic acid dissociation constants in seawater as a function of temperature and salinity. *Deep-Sea Res.* **36**, 1635–1654 (1989).
- Roy, R. N. *et al.* Determination of the ionization constants of carbonic acid in seawater. *Mar. Chem.* **44**, 249–268 (1993).

Acknowledgements. We thank T. Takahashi and D. W. Chipman for their assistance and support in measuring $p\text{CO}_2$; the crew of RV *Weatherbird II*; J. Benson, F. Howse, B. Bjork, M. Church, S. Becker and R. Little for their shipboard assistance; D. Frye and T. D. Dickey for wind data from the Bermuda Testbed Mooring Program; D. Ollis and R. Parkes for wind and air temperature data from the former US Naval Station and Bermuda Harbour Radio, respectively; N. B. Nelson and R. J. Johnson for satellite and CTD data; F. Bahr for processing thermosalinograph data; and D. A. Hansell and C. A. Carlson for their reviews. This research was supported by NSF.

Correspondence and requests for materials should be addressed to N.R.B. (e-mail: nick@bbsr.edu).

Seismic evidence for a lower-mantle origin of the Iceland plume

Yang Shen*, Sean C. Solomon†, Ingi Th. Bjarnason‡ & Cecily J. Wolfe*

* Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

† Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington DC 20015, USA

‡ Science Institute, University of Iceland, Reykjavik, Iceland

Iceland, one of the most thoroughly investigated hotspots^{1–3}, is generally accepted to be the manifestation of an upwelling mantle plume⁴. Yet whether the plume originates from the lower mantle or from a convective instability at a thermal boundary layer between the upper and lower mantle near 660 km depth^{5,6} remains unconstrained. Tomographic inversions of body-wave delay times show that low seismic velocities extend to at least 400 km depth beneath central Iceland^{7,8}, but cannot resolve structure at greater depth. Here we report lateral variations in the depths of compressional-to-shear wave conversions at the two seismic discontinuities marking the top and bottom of the mantle transition zone beneath Iceland. We find that the transition zone is 20 km thinner than in the average Earth⁹ beneath central and southern Iceland, but is of normal thickness beneath surrounding areas, a result indicative of a hot and narrow plume originating from the lower mantle.

Observational constraints on the deep structure of plumes are sparse, with only one report of the detection of a narrow plume conduit at a depth of ~700 km (ref. 10). A new approach to address the question of the depth of origin of mantle plumes is to map the mantle seismic discontinuities near 410 and 660 km depth, features that have been identified with the transitions from the α -phase to

the β -phase of $(\text{Mg,Fe})_2\text{SiO}_4$ and from γ - $(\text{Mg,Fe})_2\text{SiO}_4$ to perovskite plus magnesiowustite¹¹, respectively. The depths to the 410- and 660-km discontinuities respectively increase and decrease with increasing temperature, and thus provide information on lateral temperature variations and associated mantle circulation patterns. As sketched in Fig. 1, lateral variations in the discontinuity depths can be diagnostic of the depth of origin of mantle plumes.

The data used here are receiver functions¹² derived from body-wave records of teleseismic earthquakes from the broadband ICE-MELT seismic network¹³ and the permanent Global Seismographic Network (GSN) station BORG on Iceland (Fig. 2). The calculation of receiver functions follows procedures previously described¹⁴.

To image lateral variations in seismic discontinuities, we use geographical binning of receiver functions¹⁵. At a given binning depth beneath Iceland and the surrounding area, we divide the horizontal plane into overlapping square patches (200×200 km, Fig. 2), comparable in dimension to the Fresnel zone of a P660s phase (Pds denotes a P -to- S conversion at depth d) of frequency 0.1–0.3 Hz at the conversion depth. Each patch overlaps two-thirds of its area with adjacent patches. For every 10-km increment in binning depth from the surface to 1,200 km, receiver functions having Pds paths that pierce the same patch are gathered (Fig. 2), corrected for moveout to a reference ray parameter of 0.0573 s km^{-1} (ref. 14), and stacked using an n th-root method¹⁶ ($n = 2$) in the time window corresponding to the given binning depth interval.

The stacked receiver functions clearly reveal arrivals corresponding to the 410- and 660-km discontinuities (Fig. 3). There is no evidence for a coherent P -to- S conversion from near 520 km depth¹⁷. North of Iceland, variations in P410s and P660s times are positively correlated and comparable in magnitude (Fig. 3b). Such a positive correlation reflects velocity heterogeneities shallower than the 410-km discontinuity because of the nearly identical paths of P660s and P410s over that depth interval. P660s–P410s differential times, which provide information on the thickness of the transition zone, are not sensitive to heterogeneities shallower than the 410-km discontinuity and are similar to that predicted by the iasp91 global model⁹ (Fig. 3b). The positive correlation between P410s and P660s times breaks down beneath central and southern Iceland (Figs 3c

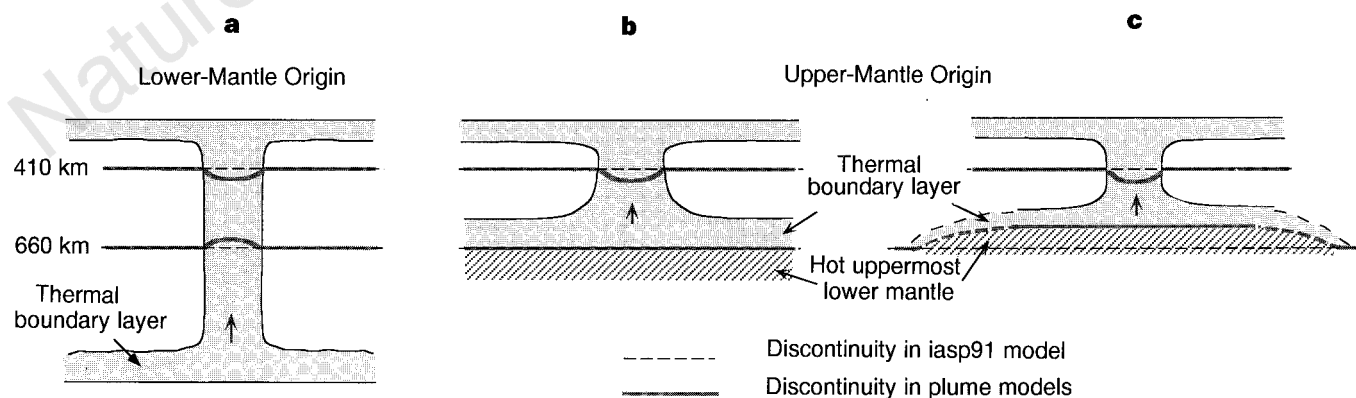


Figure 1 Variations in the depths of the 410- and 660-km phase boundaries can distinguish three generalized, end-member models for mantle plumes, here depicted schematically. Horizontal dashed lines denote 410 and 660 km depth. The seismic discontinuities associated with the phase change are shown as thick lines. The shaded regions represent hotter-than-normal mantle material associated with plumes; the stippled regions represent hot lower-mantle material. **a**, Under the model in which a plume originates from the lower mantle⁴, the depths to the 410- and 660-km discontinuities are deeper and shallower than normal, respectively, but only in the immediate vicinity of the plume conduit. **b**, In one type of strongly layered mantle convection model, plumes arise from instabilities in a global thermal boundary layer at the base of the upper mantle marked by the 660-

km phase boundary, and lateral variations in temperatures at the base of the boundary layer are generally less than the excess temperature of the plumes^{5,19}. The depth to the 660-km discontinuity is thus nearly constant and, by definition, that of the average Earth⁹. The 410-km discontinuity is deeper than normal in the vicinity of the plume conduit. **c**, In another type of layered-mantle convection model, plumes arise from regions of the boundary layer at the base of the upper mantle that are heated by the underlying lower mantle²⁰. The horizontal extent of the portion of the boundary layer supplying the plume is much greater than the diameter of the plume conduit, reducing the depth of the 660-km discontinuity, and thus the thickness of the mantle transition zone, over a broad region.

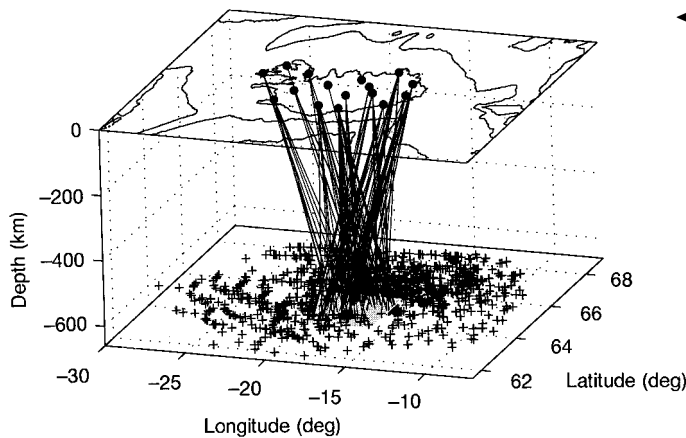


Figure 2 The geometry of our experiment. Shown are the locations of the piercing points of P660s ray paths at the 660-km discontinuity (crosses) and the P660s ray paths (lines) from the central patch (grey square) to the ICEMELT seismic stations and the GSN station BORG (dots). The Icelandic coast and the Reykjanes and Kolbeinsey ridges are represented by regional bathymetry at 1-km contours. A total of 1,560 pairs of radial and transverse receiver functions sample the discontinuity.

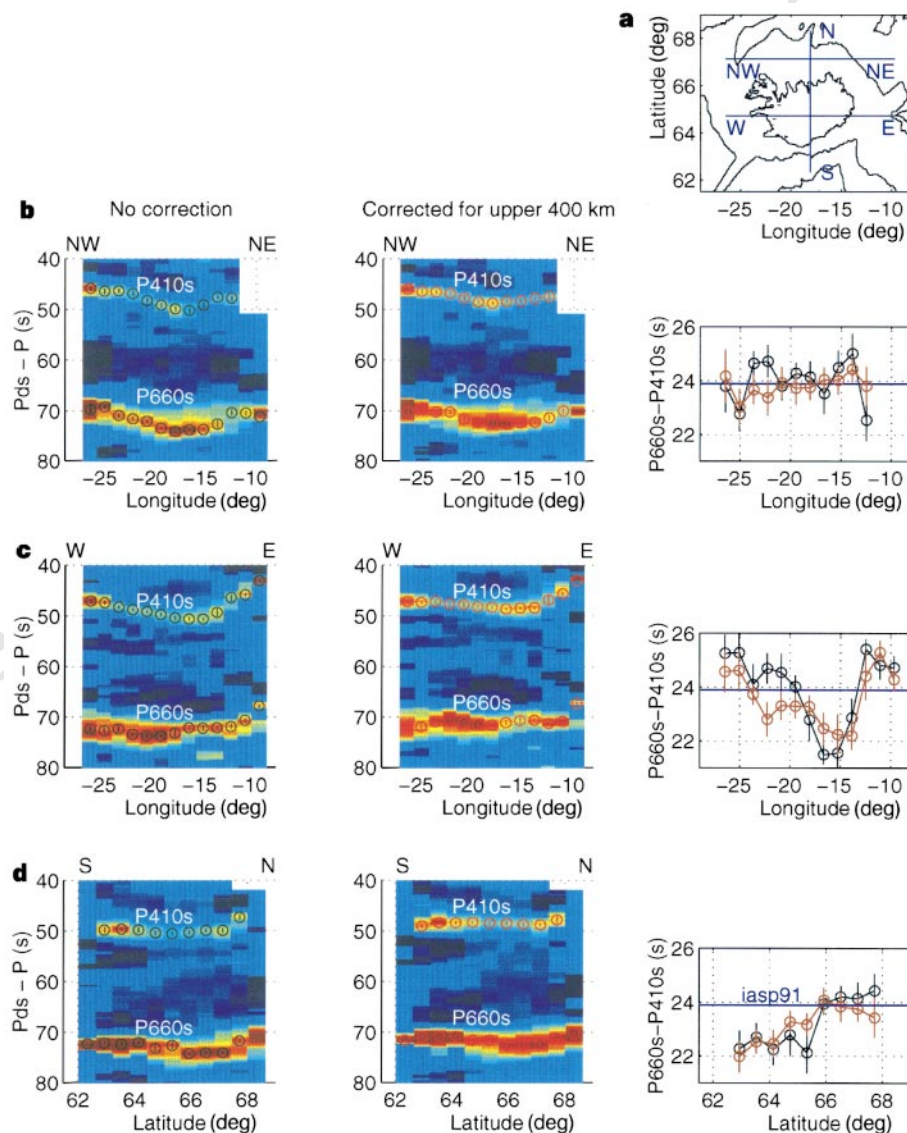


Figure 3 Images of P410s and P660s in stacked receiver functions and P660s-P410s differential times indicate an anomalously thin transition zone beneath central and southern Iceland. **a**, Locations of the profiles of receiver function stacks. **b**, The left-hand and middle panels display relative amplitudes of receiver function stacks along an east-west profile north of Iceland. Red and yellow colours represent positive and relatively higher-amplitude arrivals. We note that the n -th-root stacking¹⁶ amplifies weak signals in noisy data, but does not preserve amplitudes and waveforms. The vertical axis is the time after the compressional wave (P) arrival. Receiver functions are uncorrected (left panel) or have been corrected (middle panel) for the splitting of the converted phases and velocity

heterogeneity in the upper 400 km. P660s-P410s differential times (right panel) are similar to that predicted for the iasp91 global model⁹ (23.9 s, blue horizontal line) for both corrected (red circles) and uncorrected (black circles) receiver functions. The P410s and P660s times, the differential times, and their 1σ errors are estimated using a bootstrap method³⁰. The relative amplitudes of identified P410s and P660s phases are greater than twice the levels of similarly processed and stacked records of noise immediately before the corresponding P arrivals. A 1-s change in differential time is equivalent to about a 10-km change in the transition zone thickness. **c**, An east-west profile through central Iceland. **d**, A north-south profile through central Iceland.

and d), where observed P660s-P410s differential times are less than that predicted by iasp91 by as much as 2.2 s. We applied corrections for velocity heterogeneity in the upper 400 km using P and S tomographic models⁸ and for Pds shear-wave splitting using the results from SKS splitting analysis¹⁸. Station terms in the tomographic models⁸ are incorporated, under the assumption that they reflect a uniform contribution from the upper 100 km. These corrections do not change the general pattern of differential times and only partially remove the effects of velocity heterogeneity (Fig. 3b), because the tomographic inversions underestimate the magnitude of the velocity anomalies⁸. Stacks of randomly selected receiver functions (see Supplementary Information) show significantly less coherence for the converted phases than the correctly binned stacks, indicating that the observed discontinuity structure is real.

The mantle transition-zone thickness beneath central and southern Iceland implied by the P660s-P410s differential times is less than beneath surrounding areas by ~20 km. The location and east–west dimension (400 km) of the region of the thinner (and therefore hotter than normal) transition zone (Fig. 4) are generally consistent with the location and diameter of the low-velocity anomalies at 300 km depth imaged in the tomographic models⁸. The mantle transition-zone thickness beneath the areas surrounding central and southern Iceland is comparable to that of the iasp91 model⁹. Observations of long-period SS precursors¹⁷ indicate that the mantle transition zone in most oceanic areas is unlikely to be significantly thicker than in iasp91. This inference provides a basis for a discussion of the depth of origin of the Iceland plume.

The combination of a thinner transition zone beneath central and southern Iceland and a normal transition zone beneath surrounding areas is contrary to models in which the plume originates from an instability in a boundary layer at the base of the upper mantle as a result of regional heating from below^{19,20} (Fig. 1c). Numerical and experimental studies have shown that the horizontal extent of the portion of the boundary layer supplying the thermal and mass flux at a plume is much greater than the diameter of the plume conduit²¹; in particular, the area of a conductive boundary layer 50–100 km thick supplying the plume would have to be 1,000–2,000 km in horizontal extent to match the excess heat flux at Iceland²². A broad

boundary-layer instability (Fig. 1c) having an excess temperature of 150–200 K (ref. 2) would reduce the depth of the 660-km discontinuity and the transition-zone thickness by 8–10 km (0.8–1.0 s differential P660s-P410s time) over a broad region, inconsistent with the normal transition-zone thickness observed beneath the areas surrounding central and southern Iceland.

The thinner transition zone appears to be circular within the mapped area and does not follow the geometry of the Mid-Atlantic Ridge (Fig. 4). This pattern suggests that a thinner transition zone is not a general feature of mid-ocean ridges. In support of this observation, receiver functions from sea-floor seismic stations on the southern East Pacific Rise²³ show that the average thickness of the underlying mantle transition zone is comparable to iasp91. Precursors to SS waves reflected beneath the northern East Pacific Rise show no significant difference between the transition-zone discontinuity depths beneath sea floor younger than 10 Myr and those beneath older sea floor²⁴. Although the transition-zone thickness beneath the area southeast of Iceland remains to be mapped and we cannot fully rule out the possibility of a larger, regional anomaly, the fact that the thinner transition zone underlies the column of low seismic velocities in the tomographic models⁸ and does not follow the ridge geometry suggests that the thinner transition zone is associated with the Iceland mantle plume.

Several lines of evidence, including tomographic images of subducting slabs penetrating the lower mantle in some areas²⁵ and the localized depression of the 660-km discontinuity beneath subduction zones^{26,27}, argue against a global thermal boundary layer above the 660-km discontinuity. Under two-layer mantle convection models, upwelling beneath spreading ridges would entrain anomalously hot material from such a boundary layer²⁸ and result in a deeper than normal 410-km discontinuity and a thinner than normal transition zone in the upwelling column, contrary to a normal mantle transition-zone thickness beneath most ridges. For strongly layered mantle convection models in which the depth of the 660-km discontinuity is nearly constant^{5,19} (Fig. 1b), the low seismic velocities in the upper 400 km (ref. 8) and the thin and anomalously hot mantle transition zone beneath central and southern Iceland would result in greater delays to P660s arrivals in those areas; this is contrary to observations that P660s times beneath central and southern Iceland are less than beneath adjacent areas, except near the eastern and northern edges of the mapped region (Fig. 3c and d) where relatively greater average upper-mantle velocities appear to be present along the paths of the converted phases. Although the imaged discontinuity topography is affected either by no correction or by an undercorrection for velocity heterogeneity, corrections using seismic models with a narrower mantle plume and a larger velocity anomaly⁸—and including delays expected from higher than normal temperatures within the portion of the plume in the transition zone beneath central and southern Iceland (~0.2 s per 100 K excess temperature)—would only further shoal the apparent depth of the 660-km discontinuity and reduce the transition-zone thickness beneath central and southern Iceland.

Our observations are therefore most consistent with a lower-mantle origin for the Iceland plume (Fig. 1a). The variation in the depth to the 660-km discontinuity beneath Iceland is also consistent with our premise that the observed discontinuity corresponds to a mineralogical phase boundary, rather than a chemical boundary, because a chemically stratified boundary between the upper and lower mantle would be deflected by as much as several hundred kilometres by an upwelling plume^{19,20} from the lower mantle. For Clapeyron slopes of 2.9 and –2.1 MPa K^{–1} for the 410- and 660-km discontinuities²⁹, respectively, the reduction of the transition-zone thickness by ~20 km beneath central and southern Iceland is equivalent to an excess temperature of 150 K, a result in good agreement with estimates of the thermal anomaly at the depth of melt generation (<200 km)². The inferred excess temperature

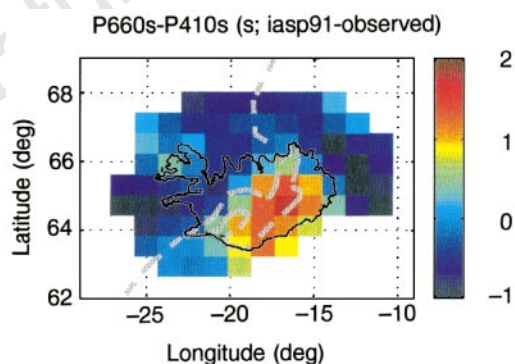


Figure 4 Map view of differences between the observed P660s-P410s differential times and the value predicted for the iasp91 model⁹ underscores the localized nature of the anomaly. Receiver functions have not been corrected for shear-wave splitting or velocity heterogeneity in the upper 400 km; such corrections do not change the general pattern of differential times (Fig. 3 and Supplementary Information). Red and yellow colours indicate significantly smaller differential times (thinner mantle transition zone) than in iasp91, while blue colours denote normal or somewhat greater differential times (normal or slightly thicker mantle transition zone). The images have been smoothed by a two-dimensional, five-point moving average. The dashed grey lines delineate Icelandic neovolcanic zones and the axis of the Mid-Atlantic Ridge.

(150 K) and radius of the plume (200 km) should be regarded as lower and upper bounds, respectively, because the finite sizes of both the Fresnel zones of the converted phases and the patches used for stacking tend to smooth lateral variations in discontinuity depths (see Supplementary Information). These values nonetheless provide strong support for models of a hot and narrow plume penetrating the mantle transition zone beneath Iceland, and are similarly strong evidence against models in which the plume is broader (radius > 300 km) and has less excess temperature ($\Delta T \approx 70$ K)³. □

Received 9 September 1997; accepted 1 June 1998.

- Schilling, J. G. Iceland mantle plume: Geochemical study of Reykjanes Ridge. *Nature* **242**, 565–571 (1973).
- White, R. S., Brown, J. W. & Smallwood, J. R. The temperature of the Iceland plume and the origin of outward-propagating V-shaped ridges. *J. Geol. Soc. Lond.* **152**, 1039–1045 (1995).
- Ribe, N. M., Christensen, U. R. & Theissing, J. The dynamics of plume-ridge interaction, 1: Ridge-centered plumes. *Earth Planet. Sci. Lett.* **134**, 155–168 (1995).
- Morgan, W. J. Convection plumes in the lower mantle. *Nature* **230**, 42–43 (1971).
- Richter, F. M. & McKenzie, D. P. On some consequences and possible causes of layered mantle convection. *J. Geophys. Res.* **86**, 6133–6142 (1981).
- Anderson, D. L. Hotspots, basalts and the evolution of the mantle. *Science* **213**, 82–89 (1981).
- Tryggvason, K., Husebye, E. S. & Stefánsson, R. Seismic image of the hypothesized Icelandic hot spot. *Tectonophysics* **100**, 94–118 (1983).
- Wolfe, C. J., Bjarnason, I. Th., VanDecar, J. C. & Solomon, S. C. Seismic structure of the Iceland mantle plume. *Nature* **385**, 245–247 (1997).
- Kennett, B. L. N. & Engdahl, E. R. Travel times for global earthquake location and phase identification. *Geophys. J. Int.* **105**, 429–466 (1991).
- Nataf, H.-C. & VanDecar, J. C. Seismological detection of a mantle plume? *Nature* **364**, 115–120 (1993).
- Liu, L.-G. in *Earth: Its Origin, Structure and Evolution* (ed. McElhinny, M. W.) 177–202 (Academic, San Diego, 1979).
- Langston, C. A. Structure under Mount Rainier, Washington, inferred from teleseismic body waves. *J. Geophys. Res.* **84**, 4749–4762 (1979).
- Bjarnason, I. Th., Wolfe, C. J., Solomon, S. C. & Gudmundson, G. Initial results from the ICEMELT experiment: Body-wave delay times and shear-wave splitting across Iceland. *Geophys. Res. Lett.* **23**, 459–462 (1996); Correction. *Geophys. Res. Lett.* **23**, 903 (1996).
- Shen, Y., Solomon, S. C., Bjarnason, I. Th. & Purdy, G. M. Hot mantle transition zone beneath Iceland and the adjacent Mid-Atlantic Ridge inferred from P-to-S conversions at the 410- and 660-km discontinuities. *Geophys. Res. Lett.* **23**, 3527–3530 (1996).
- Dueker, K. G. & Sheehan, A. F. Mantle discontinuity structure from midpoint stacks of converted P to S waves across the Yellowstone hotspot track. *J. Geophys. Res.* **102**, 8313–8327 (1997).
- Kanasewich, E. R., Hemmings, C. D. & Alpaslan, T. Nth-root stack nonlinear multichannel filter. *Geophysics* **38**, 327–338 (1973).
- Shearer, P. M. Transition zone gradients and the 520-km discontinuity. *J. Geophys. Res.* **101**, 3053–3066 (1996).
- Bjarnason, I. Th., Silver, P. G. & Solomon, S. C. Teleseismic shear wave splitting near the Iceland hotspot: Results from ICEMELT (abstr.). *Eos* **78**, S322 (1997).
- Nakakuki, T., Sato, H. & Fujimoto, H. Interaction of the upwelling plume with the phase and chemical boundary at the 670 km discontinuity: Effects of temperature-dependent viscosity. *Earth Planet. Sci. Lett.* **121**, 369–384 (1994).
- Kellogg, L. H. Interactions of plumes with a compositional boundary at 670 km. *Geophys. Res. Lett.* **18**, 865–868 (1991).
- Olson, P., Schubert, G., Anderson, C. & Goldman, P. Plume formation and lithosphere erosion: A comparison of laboratory and numerical experiments. *J. Geophys. Res.* **93**, 15065–15084 (1988).
- Sleep, N. H. Hotspots and mantle plumes: Some phenomenology. *J. Geophys. Res.* **95**, 6715–6736 (1990).
- Shen, Y., Sheehan, A. F., Dueker, K. G., de Groot-Hedlin, C. & Gilbert, H. Mantle discontinuity structure beneath the southern East Pacific Rise from P-to-S converted phases. *Science* **280**, 1232–1235 (1998).
- Lee, D.-K. & Grand, S. P. Depths of the upper mantle discontinuities beneath the East Pacific Rise. *Geophys. Res. Lett.* **23**, 3369–3372 (1996).
- van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
- Wicks, C. W. & Richards, M. A. A detailed map of the 660-km discontinuity beneath the Izu-Bonin subduction zone. *Science* **261**, 1424–1427 (1993).
- Niu, F. & Kawakatsu, H. Direct evidence for the undulation of the 660-km discontinuity beneath Tonga: Comparison of Japan and California array data. *Geophys. Res. Lett.* **22**, 531–534 (1995).
- Zhong, S. & Gurnis, M. Role of plates and temperature-dependent viscosity in phase change dynamics. *J. Geophys. Res.* **99**, 15903–15917 (1994).
- Bina, C. R. & Helffrich, G. Phase transition Clapeyron slopes and transition zone seismic discontinuity topography. *J. Geophys. Res.* **99**, 15853–15860 (1994).
- Efron, B. & Gong, G. A leisure look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.* **37**, 36–48 (1983).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank Björn Bjarnason, Birgir Bjarnason, B. Brandsdóttir, H. Brynjólfsson, K. Egilsson, G. Gudmundson, E. Hannesson, T. Hardarson, L. Helgason, B. Ingimundarson, H. Jónsson, E. Kjartansson, A. Kuehnel, R. Kuehnel, P. Sigurdsson, R. Thordarsson, and the staff of the National Electric Company of Iceland (Landsvirkjun) for assistance with field operations; the IRIS-DMS for making available data from GSN station BORG; G. M. Purdy, D. W. Forsyth, R. S. Detrick, and J. A. Collins for encouragement and advice in the early stages of this work; S. van der Lee for providing software; D. E. James, E. M. Parmentier, I. S. Sacks, and P. G. Silver for discussions, and G. Helffrich and P. van Keken for reviews. This work was supported by the US NSF.

Correspondence and requests for materials should be addressed to Y.S. (e-mail: yshen@whoi.edu).

Microbiological evidence for Fe(III) reduction on early Earth

Madeline Vargas*, Kazem Kashefi, Elizabeth L. Blunt-Harris & Derek R. Lovley

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts 01003, USA

It is generally considered¹ that sulphur reduction was one of the earliest forms of microbial respiration, because the known microorganisms that are most closely related to the last common ancestor of modern life are primarily anaerobic, sulphur-reducing hyperthermophiles^{2–4}. However, geochemical evidence indicates that Fe(III) is more likely than sulphur to have been the first external electron acceptor of global significance in microbial metabolism^{5–7}. Here we show that Archaea and Bacteria that are most closely related to the last common ancestor can reduce Fe(III) to Fe(II) and conserve energy to support growth from this respiration. Surprisingly, even *Thermotoga maritima*, previously considered to have only a fermentative metabolism, could grow as a respiratory organism when Fe(III) was provided as an electron acceptor. These results provide microbiological evidence that Fe(III) reduction could have been an important process on early Earth and suggest that microorganisms might contribute to Fe(III) reduction in modern hot biospheres. Furthermore, our discovery that hyperthermophiles that had previously been thought to require sulphur for cultivation can instead be grown without the production of toxic and corrosive sulphide, should aid biochemical investigations of these poorly understood organisms.

Our understanding of the geochemical conditions at the time when respiratory systems were first evolving suggests that the number of potential electron acceptors for microbial respiration was much more limited than at present. Important modern electron acceptors such as oxygen, nitrate and sulphate are unlikely to have been present in quantities sufficient to support globally significant rates of respiration^{8,9}. However, Fe(III), derived from photochemical oxidation of Fe(II) in Archaean seas and from hydrothermal vent fluids, is thought to have been abundant on early Earth^{5–7,10}. To examine whether Fe(III) reduction might have been a metabolic feature of ancient anaerobic microorganisms, various hyperthermophilic microorganisms were screened for their capacity for Fe(III) reduction. The rationale for this was that geochemical and microbiological evidence indicates that early life might have evolved in or near hydrothermal zones and thus extant hyperthermophilic microorganisms that are most closely related to the last common ancestor (LCA) may serve as models for early forms of microbial respiration^{2,5,11,12}. Although a capacity for Fe(III) reduction has been found in a variety of microorganisms, including some moderate thermophiles^{13–18}, none of these is closely related to the LCA. Furthermore, the phylogenetic placement of these organisms in small, discrete groups does not provide convincing evidence for Fe(III) being an early metabolic characteristic¹³.

All of the hyperthermophiles were found to reduce Fe(III) (Fig. 1). In each case, Fe(III) reduction was enzymatic as there was no Fe(III) reduction in controls without cells, or when cells were incubated at 35 °C, a temperature too low for hyperthermophilic enzymatic activity. The capacity for Fe(III) reduction was constitutive, as these organisms reduced Fe(III) even though they had been grown with other electron acceptors. This is similar to mesophilic Fe(III)-reducing bacteria, which also constitutively produce Fe(III) reductase when grown anaerobically with electron acceptors other than Fe(III)¹⁴.

* Present address: Department of Biology, College of the Holy Cross, Worcester, Massachusetts 01610, USA

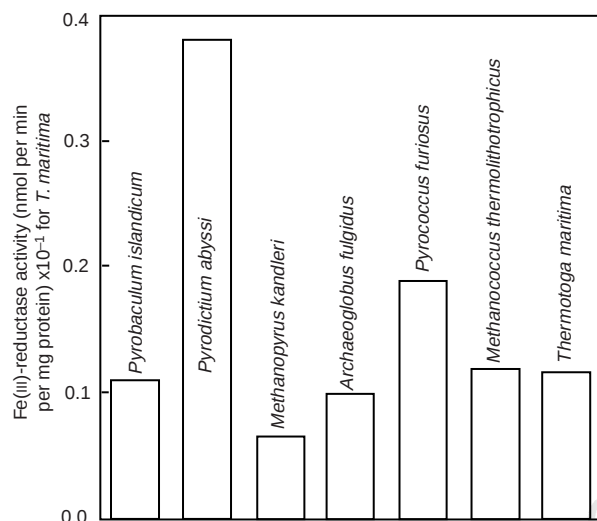


Figure 1 Rates of Fe(III) reduction in cell suspensions of hyperthermophiles.

Pyrobaculum islandicum and *Thermotoga maritima* were chosen for further study because of the generally recognized ease in culturing these organisms and because they provided examples of Archaea and Bacteria that were closely related to the LCA. Both hyperthermophiles grew in medium with H₂ as the electron donor and Fe(III) as the electron acceptor (Figs 2 and 3). There was no growth when either H₂ (Figs 2 and 3) or Fe(III) (data not shown) was omitted. Although these studies were done using soluble Fe(III), both *P. islandicum* and *T. maritima* grew readily in medium in which insoluble Fe(III) oxide was the electron acceptor, converting the Fe(III) oxide to magnetite in a manner similar to that observed previously in cultures of mesophilic Fe(III) reducers^{19,20}. This is significant because magnetite accumulations that might be attributed to Fe(III)-reducing microorganisms have been used as evidence that Fe(III) reduction was the first globally significant respiratory process on Earth⁶, as well as evidence for early life on Mars²¹, and for a hot deep terrestrial biosphere²².

The finding that a diversity of microorganisms most closely related to the LCA have the ability to reduce Fe(III) suggests that the LCA was an Fe(III)-reducing microorganism. For organisms most closely related to the LCA, Fe(III) reduction seems to be a more

universal metabolic characteristic than elemental sulphur (S⁰) reduction, because hyperthermophiles such as *Archaeoglobus fulgidus* that do not reduce S⁰ do reduce Fe(III). More significant is the finding that with Fe(III) as the electron acceptor, *T. maritima* can grow as a respiratory organism, whereas it was previously considered to have only a fermentative metabolism as it does not conserve energy from S⁰ or thiosulphate reduction^{3,4,23,24}. This suggests that Fe(III) reduction is more closely associated with central metabolism in some hyperthermophiles than S⁰ reduction. However, the role of Fe(III) reduction in the metabolism of all of the hyperthermophiles found to have the capacity for Fe(III) reduction is not yet known. For example, the methanogens continued to produce methane during Fe(III) reduction, which makes it difficult to determine whether energy to support growth is conserved during Fe(III) reduction by these organisms. In addition to investigating Fe(III) reduction by hyperthermophiles, it will be interesting to determine whether the as-yet uncultured Archaea, whose 16S ribosomal DNA sequences have been found in cooler environments such as soils and aquatic sediments¹², can reduce Fe(III) and whether isolation with medium containing Fe(III) as the electron acceptor may be a productive strategy for culturing these organisms.

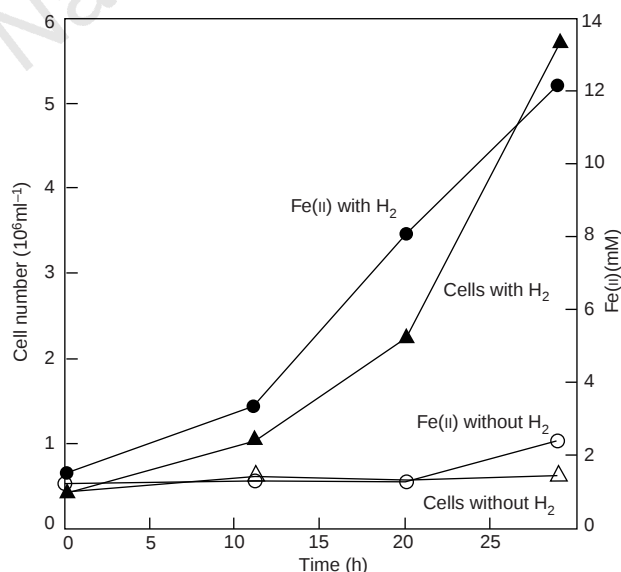


Figure 2 Growth of *Pyrobaculum islandicum* with H₂ as the electron donor and Fe(III) as the electron acceptor.

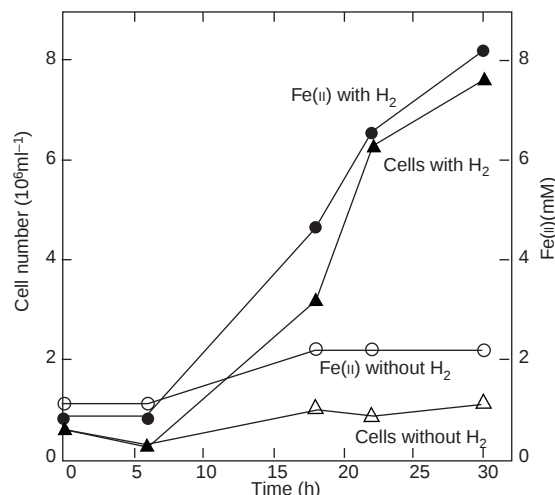


Figure 3 Growth of *Thermotoga maritima* with H₂ as the electron donor and Fe(III) as the electron acceptor.

Although the LCA was likely to have been a metabolically sophisticated respiratory organism^{2,5,12}, earlier more primitive microorganisms might also have had the capacity to transfer electrons to extracellular Fe(III). The evolution of this metabolic capability would have provided organisms with an energetic advantage over early fermentative microorganisms by greatly expanding the range and extent of potential electron-donor oxidation. Once reduced, iron can readily transfer electrons to other electron acceptors, and this may have led to the intracellular incorporation of Fe(III)/Fe(II) redox couples that are common in more complex, modern electron-transport pathways. □

Methods

Organisms were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ). For studies with cell suspensions, cells were grown in the medium suggested by the DSMZ for their cultivation. Cells were collected by centrifugation under N₂-CO₂ (80:20), and washed cell suspensions were prepared in anaerobic bicarbonate buffer (30 mM; pH 6.8) to provide ~0.1 mg of cell protein in 5 ml bicarbonate buffer under N₂-CO₂. H₂ (101 kPa) was added as the electron donor and Fe(III) citrate (5 mM) was added as the electron acceptor. For studies on the growth of *P. islandicum* with Fe(III), cells were grown in DSMZ medium 390 in which thiosulphate and sodium sulphide had been replaced by Fe(III) citrate (15 mM) as the electron acceptor and 0.25 mM cysteine as a sulphur source, with 0.02% yeast extract added. H₂ was provided at 101 kPa; the incubation was done at 90 °C. For studies on the growth of *T. maritima* with Fe(III), cells were grown in TB medium²⁵ modified by substituting PIPES buffer with bicarbonate (30 mM), decreasing the cysteine to 0.5 mM, adding 0.05% yeast extract, and supplying H₂ (101 kPa) and Fe(III) citrate (15 mM); incubation was at 80 °C. For growth on Fe(III) oxide, organisms were grown on the media described but with Fe(III) citrate replaced with 100 mM Fe(III) oxide²⁶. Cell numbers were monitored with acridine orange staining and epifluorescent microscopy²⁷. Production of Fe(II) was quantified with ferrozine²⁶.

Received 15 April; accepted 1 July 1998.

- Madigan, M. T., Martinko, J. M. & Parker, J. *Biology of Microorganisms* (Prentice Hall, Upper Saddle River, New Jersey, 1997).
- Pace, N. R. Origin of life—facing up to the physical setting. *Cell* **65**, 531–533 (1991).
- Adams, M. W. W. Enzymes and proteins from organisms that grow near and above 100 °C. *Annu. Rev. Microbiol.* **47**, 627–658 (1993).
- Stetter, K. O. Hyperthermophilic prokaryotes. *FEMS Microbiol. Rev.* **18**, 149–158 (1996).
- de Duve, C. *Vital Dust* 1–362 (Basic Books, New York, 1995).
- Walker, J. C. G. Was the Archaean biosphere upside down? *Nature* **329**, 710–712 (1987).
- Cairns-Smith, A. G., Hall, A. J. & Russell, M. J. Mineral theories of the origin of life and an iron sulfide example. *Orig. Life Evol. Biosph.* **22**, 161–180 (1992).
- Walker, J. C. G. & Brimblecombe, P. Iron and sulfur in the pre-biologic ocean. *Precamb. Res.* **28**, 205–222 (1985).
- Kasting, J. F. Earth's early atmosphere. *Science* **259**, 920–926 (1993).
- De Ronde, C. E. J., De Wit, M. J. & Spooner, E. T. C. Early Archean (<3.2 Ga) Fe-oxide-rich, hydrothermal discharge vents in the Barberton greenstone belt, South Africa. *Geol. Soc. Am. Bull.* **106**, 86–104 (1984).
- Holm, N. G. Why are hydrothermal systems proposed as plausible environments for the origin of life? *Origins Life Evol. Biosph.* **22**, 5–14 (1992).
- Barns, S. M. & Nierzwicki-Bauer, S. A. Microbial diversity in ocean, surface, and subsurface environments. *Rev. Mineral.* **35**, 35–79 (1997).
- Loneragan, D. J. *et al.* Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J. Bacteriol.* **178**, 2402–2408 (1996).
- Lovley, D. R., Coates, J. D., Saffarini, D. A. & Loneragan, D. J. in *Iron and Related Transition Metals in Microbial Metabolism* (eds Winkelman, G. & Carrano, C. J.) 187–215 (Harwood, Switzerland, 1997).
- Boone, D. R. *et al.* *Bacillus infernus* sp. nov., an Fe(III)- and Mn(IV)-reducing anaerobe from the deep terrestrial subsurface. *Int. J. Sys. Bacteriol.* **45**, 441–448 (1995).
- Slobodkin, A., Reysenbach, A.-L., Strutz, N., Dreier, M. & Wiegand, J. *Thermoterrabacterium ferrirubens* gen. nov., sp. nov., a thermophilic anaerobic dissimilatory Fe(III)-reducing bacterium from a continental hot spring. *Int. J. System. Bacteriol.* **47**, 541–547 (1997).
- Greene, A. C., Patel, B. K. C. & Sheehy, A. J. *Deferribacter thermophilus* gen. nov., sp. nov., a novel thermophilic manganese- and iron-reducing bacterium isolated from a petroleum reservoir. *Int. J. System. Bacteriol.* **47**, 505–509 (1997).
- Liu, S. V. *et al.* Thermophilic Fe(III)-reducing bacteria from the deep subsurface: the evolutionary implications. *Science* **277**, 1106–1109 (1997).
- Lovley, D. R., Stolz, J. F., Nord, G. L. & Phillips, E. J. P. Anaerobic production of magnetite by a dissimilatory iron-reducing microorganism. *Nature* **330**, 252–254 (1987).
- Lovley, D. R. in *Iron Biominerals* (eds Frankel, R. B. & Blakemore, R. P.) 151–166 (Plenum, New York, 1990).
- McKay, D. S. *et al.* Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
- Gold, T. The deep, hot biosphere. *Proc. Natl Acad. Sci. USA* **89**, 6045–6049 (1992).
- Huber, R. *et al.* *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °C. *Arch. Microbiol.* **144**, 324–333 (1986).
- Ravot, G. *et al.* Thiosulfate reduction, an important physiological feature shared by members of the order *Thermotogales*. *Appl. Environ. Microbiol.* **61**, 2053–2055 (1995).
- Childers, S., Vargas, M. & Noll, K. M. Improved methods for cultivation of the extremely thermophilic bacterium *Thermotoga neapolitana*. *Appl. Environ. Microbiol.* **58**, 3949–3953 (1992).
- Lovley, D. R. & Phillips, E. J. P. Organic matter mineralization with reduction of ferric iron in anaerobic sediments. *Appl. Environ. Microbiol.* **51**, 683–689 (1986).
- Hobbie, J. E., Daley, R. J. & Jasper, S. Use of Nuclepore filters for counting bacteria by fluorescence microscopy. *Appl. Environ. Microbiol.* **33**, 1225–1228 (1977).

Acknowledgements. This work was supported by a grant from the LExEN program of the NSF.

Correspondence and requests for materials should be addressed to D.R.L. (e-mail: dlovley@microbio.umass.edu).

The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird

Michale S. Fee*, Boris Shraiman*, Bijan Pesaran† & Partha P. Mitra*

* Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA

† Physics Department, California Institute of Technology, Pasadena, California 91125, USA

Birdsong is characterized by the modulation of sound properties over a wide range of timescales¹. Understanding the mechanisms by which the brain organizes this complex temporal behaviour is a central motivation in the study of the song control and learning system^{2–8}. Here we present evidence that, in addition to central neural control, a further level of temporal organization is provided by nonlinear oscillatory dynamics that are intrinsic to the avian vocal organ. A detailed temporal and spectral examination of song of the zebra finch (*Taeniopygia guttata*) reveals a class of rapid song modulations that are consistent with transitions in the dynamical state of the syrinx. Furthermore, *in vitro* experiments show that the syrinx can produce a sequence of oscillatory states that are both spectrally and temporally complex in response to the slow variation of respiratory or syringeal parameters. As a consequence, simple variations in a small number of neural signals can result in a complex acoustic sequence.

Zebra-finch songs are organized into short segments of sound, referred to as syllables, that often contain rapid sequences of whistled notes, harmonic stacks and aperiodic signals^{9,10}. Sounds within a syllable are produced contiguously with no silent interval during the transitions between distinct signal types (Fig. 1a). This continuity allows one to observe changes in the oscillatory state of the syrinx during modulations of the acoustic signal. We recorded the directed song of male zebra finches (*n* = 12) in a sound-isolation chamber. Close examination of the song spectra and acoustic waveforms shows sudden transitions from periodic to aperiodic or chaotic dynamics, period doubling, and mode-locking transitions. We believe that these features arise because the syrinx behaves as a low-dimensional nonlinear dynamical system.

Most (10 of 12) of the songs we recorded contained syllables exhibiting rapid (<10 ms) transitions from periodic to chaotic, or noisy, signals. Figure 1b shows two examples of this behaviour at an expanded timescale. These transitions sometimes occur in less than 1 ms, and often appear to jump back and forth between the periodic and chaotic oscillation. Another characteristic of nonlinear dynamical systems, period doubling, is characterized by a change in oscillation frequency such that spectral components appear at half the original frequency spacing. An example of period doubling in zebra-finch song is shown in Fig. 1c; at the time marked with the asterisk (*), note the rapid onset of interspersed harmonic components as described above. Period-doubling transitions were seen in three of the zebra-finch songs examined.

Zebra-finch song syllables can also exhibit modulations that suggest the presence of mode locking in the syringeal dynamics.

Mode locking occurs when two components of a system are constrained by some nonlinear interaction to maintain a small integer ratio of oscillation frequencies. If the characteristic frequency of one component of the system is changed relative to the other, mode-locking transitions may occur as the oscillation frequency suddenly jumps to achieve a new stable integer ratio with the

fixed frequency^{11,12}. In zebra-finch song, for example, syllables often exhibit¹⁰ an overall downward sweep in fundamental frequency. However, this downward frequency sweep may be punctuated by steps, or rapid transitions in fundamental frequency (Fig. 1c). In the example shown, the mode-locking transitions occur such that the oscillation frequency maintains a ratio of two, three and four with respect to a common frequency at ~ 3 kHz. Similar features were observed in 9 of the 12 zebra-finch songs studied.

We suggest two hypotheses for the origin of the song features described above: direct muscular manipulation during singing of the syringeal parameters by the song control system, or alternatively, intrinsic dynamics of the syrinx. To distinguish between these hypotheses, we examined the acoustic signals generated by an *in vitro* preparation of the zebra finch syrinx ($n = 6$). The excised syrinx was induced to oscillate by applying suction at the trachea and pressure at the bronchial inlet. *In vitro* oscillations of the zebra-finch syrinx were quite similar in oscillation frequency and amplitude to normal zebra-finch song and occurred over a range of pressures and flow rates that was physiologically reasonable¹³. Spectra of the sound signals measured outside the syrinx (Fig. 2a) show an enhancement of lower harmonics compared with normal song spectra. The high harmonic components in song are thought to be produced by the rapid pressure fluctuations associated with closure of the bronchial lumen during syringeal oscillation; we expect that such rapid fluctuations would be relatively suppressed in the sound signal measured outside the syrinx.

A smooth ramp-down and ramp-up of air pressure applied at the tracheal outlet produced a sequence of oscillatory states, typically with jumps in oscillation frequency and, at higher pressures, transitions from periodic to aperiodic oscillation (Fig. 2a). The precise sequence of acoustic signals produced during this manipulation was different for each preparation. The acoustic sequence was also sensitive to small changes in various experimental parameters, such as tension on the bronchus. Intrinsic differences in syringes and sensitivity to parameters may partly account for some variations across individuals and variability within individuals, respectively. Another feature that was commonly observed was hysteresis in the acoustic signal; the sequence of acoustic signals generated as the pressure was ramped up usually differed in a repeatable manner from the sequence as the pressure was ramped down. The transitions observed during this manipulation were usually abrupt, often occurring in approximately one period of the acoustic signal (Fig. 2b).

Further evidence for nonlinear dynamics of the syrinx was seen during simple manipulations of syringeal configuration. Deflection of the labium externum (LE) (Fig. 3) into the bronchial lumen constricts the flow of air through the syrinx, and is thought to be a normal articulation during song production¹⁴. In all of the syringes tested ($n = 3$), this constriction resulted in a large increase in the fundamental oscillation frequency during *in vitro* oscillation. However, in two syringes the frequency was not a smooth function of LE displacement; rather the frequency exhibited sudden jumps at certain discrete displacements (Fig. 2c). Furthermore, the oscillation seems to be constrained to frequencies that are close to a small-integer sub-harmonic of a common (higher) frequency. In the example shown, the frequency jumps from ~ 700 Hz to $\sim 1,100$ Hz to $\sim 2,200$ Hz, which have periods, approximately, of three, two and one times the period of the common 2,200-Hz component, respectively.

To understand the origin of the apparent strong nonlinearity in the syrinx, we used a stroboscopic imaging technique to visualize *in vitro* oscillations of the zebra-finch syrinx ($n = 5$) (Fig. 3). *In vitro*, there was a threshold flow rate (~ 0.5 standard litres per minute) at which the labium internum (LI) and the medial tympaniform membrane (MTM) began to oscillate with large amplitude motion (~ 0.5 mm). The motion of these membranes was phase-locked to each other and to the generated sound, suggesting that

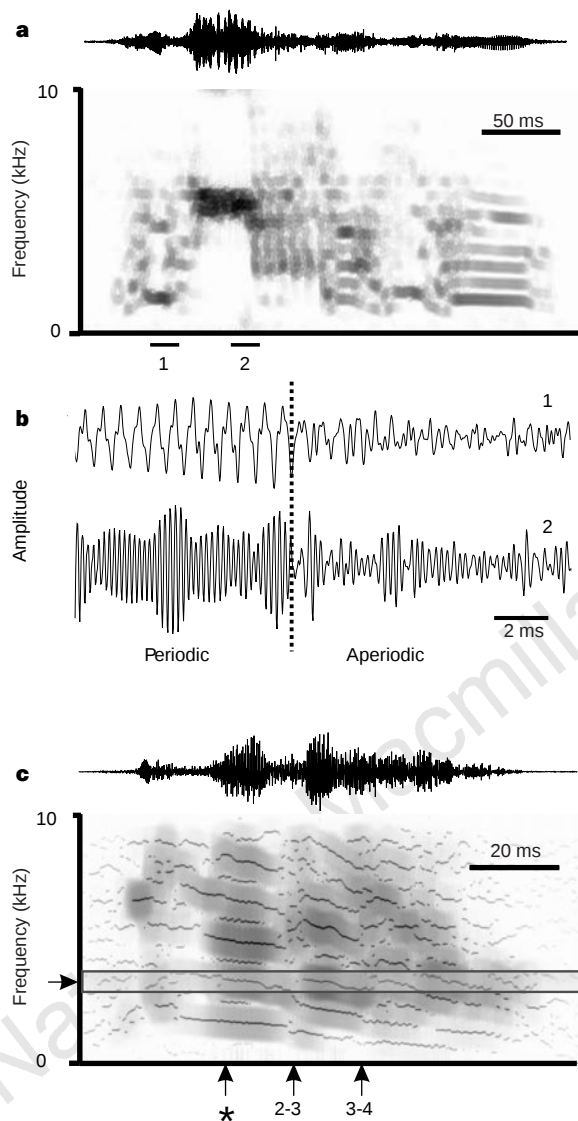


Figure 1 Spectral and temporal characteristics of zebra-finch song. **a**, The spectrogram of a syllable displaying the wide variety of spectral features common in zebra-finch song, such as a high-frequency nearly sinusoidal signal, broad-band aperiodic signals, and periodic signals characterized by harmonic stacks. The pressure waveform is shown at the top. **b**, The rapid transitions between these signals are seen by expanding the pressure waveform around the regions marked 1 and 2 in **a**. In both cases, note that the transition from a nearly periodic signal to a noisy signal occurs in roughly 1 ms. **c**, Spectrogram and harmonic analysis of a syllable recorded from a different bird. The spectrogram is shown superimposed with the result of a harmonic analysis in which sinusoids are fit to the signal in the same time window on which the spectra were computed. Note the rapid frequency-doubling transition at the time marked with the asterisk. The overall downward sweep in the fundamental frequency of this syllable is marked by a series of jumps, or sudden transitions. A rapid transition in oscillation frequency occurs at the point marked (2-3) such that the third harmonic after the transition has approximately the same frequency as the second harmonic before the transition. Approximately 15 ms later, another transition occurs such that the fourth harmonic is suddenly shifted down to the frequency of the third harmonic. The second, third and fourth harmonic in this sequence are all roughly aligned with the bright spectra band at 3,200 Hz, indicated by the horizontal bar and arrow.

these membranes are in fact the source of the sound. In all syringes studied, the anatomical structure exhibiting the largest motion amplitude was the LI, rather than the much lighter MTM, which agrees with recent evidence that the LI plays a central role in syrinx

function¹⁵. Close examination of the strobed oscillation under a microscope indicated that the bronchial lumen is probably closed briefly by contact of the LI with the LE during part of the oscillatory cycle; however, the wave-like motion of the medial membranes made it difficult to determine quantitatively the fraction of time during the oscillation that the bronchial lumen was closed. Because the excursions of these membranes are comparable to the size of the bronchial lumen, the Bernoulli forces driving the oscillation¹⁶ are a highly nonlinear function of membrane displacement.

To determine if our stroboscopic and acoustic observations of *in vitro* syringeal oscillations are consistent with a view of the syrinx as a nonlinear dynamical system, we developed a biophysical model of the syrinx. We followed previous work on both human¹⁷ and avian¹⁶ vocalizations, as well as models of fluid flow through collapsible

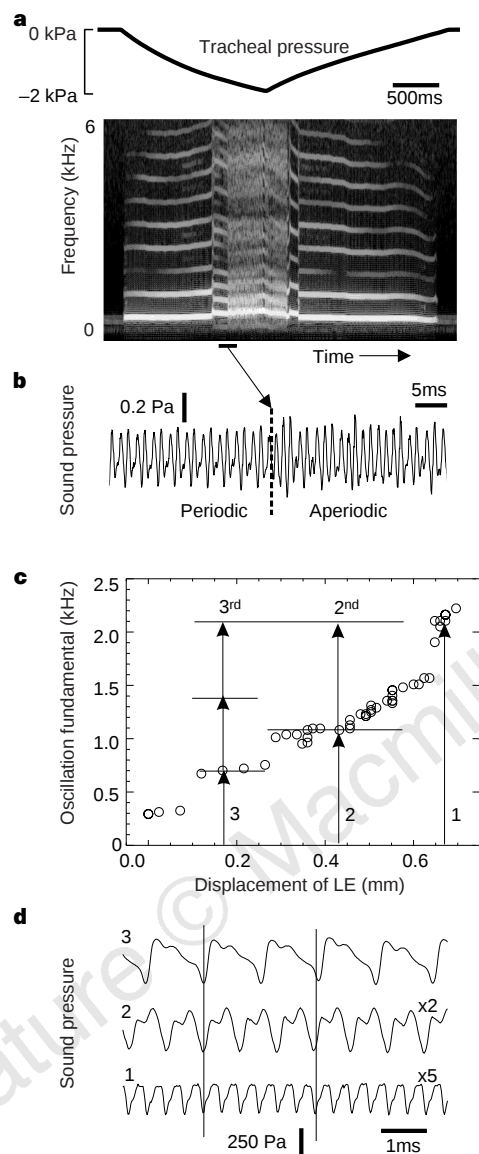


Figure 2 Patterns of acoustic signals generated in the *in vitro* syrinx preparation during slow modulation of syringeal and flow parameters. Application of suction to the trachea and slight over-pressure to one bronchus results in sustained oscillation of the ipsilateral syringeal membranes. **a**, As the tracheal pressure (top) is varied over a range of 2 kPa, the sound generated by syrinx exhibits a sequence of modulations in fundamental frequency, including period doubling, as well as a transition from periodic oscillation to broadband, apparently chaotic oscillations. Bronchial pressure is held constant (0.4 kPa). The sound is recorded outside the syrinx, ~2 cm from the MTM. **b**, Examination of the pressure signal around the transition from periodic to chaotic oscillation shows that the transition occurs in approximately one oscillatory period. **c**, Displacement of the third bronchial ring (and thus the LE) into the bronchial lumen increases the fundamental oscillation frequency over a range of nearly three octaves before the oscillation ceases. As a function of LE displacement, the fundamental frequency exhibits a series of discrete jumps. In this example, the third harmonic, second harmonic and fundamental are roughly aligned with a common frequency at ~2.2 kHz. **d**, Acoustic waveform measured at labial displacements indicated in **c**. Note the entrainment of the oscillation fundamental to three, two and one period(s) of the high-frequency component. The three waveform segments are shown at different scales, as indicated. The sound was recorded inside the trachea with a probe microphone.

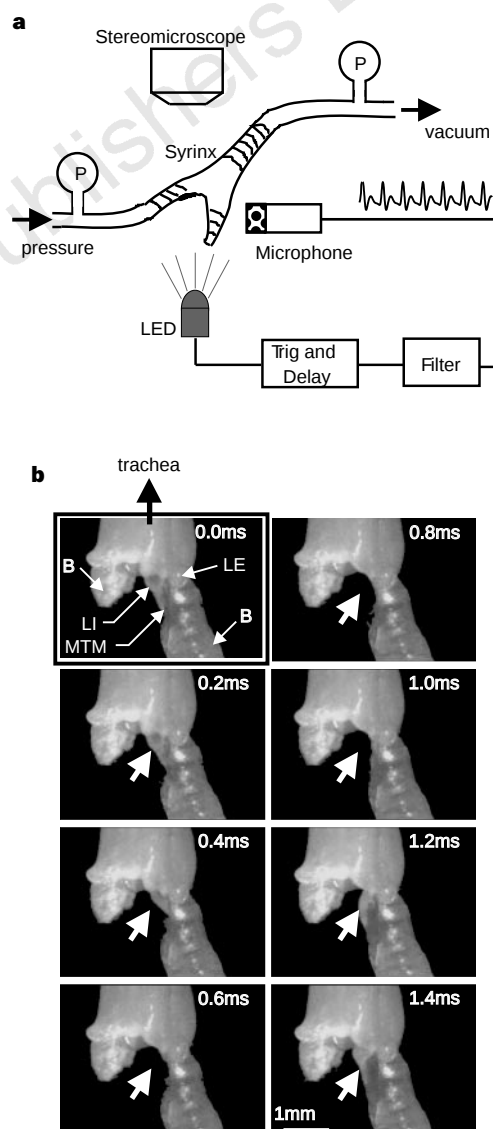


Figure 3 Stroboscopic observation of *in vitro* syringeal oscillations. **a**, Experimental setup. The syrinx is induced to oscillate *in vitro* by applying suction to the trachea and slight overpressure to one bronchial tube. The sound pressure measured ~2 cm from the syrinx is used to trigger an array of bright LEDs with a variable time delay. **b**, The configuration of the medial tympaniform membrane (MTM) and labium internum (LI) was visualized and recorded as the time delay was slowly swept through several oscillation periods. The digitized video frames are shown at increments of 0.2 ms in the time delay. In this case the oscillation frequency was ~700 Hz. Note that both the MTM and LI exhibit a large amplitude of motion. Also labelled are the labium externum (LE) and bronchi (B).

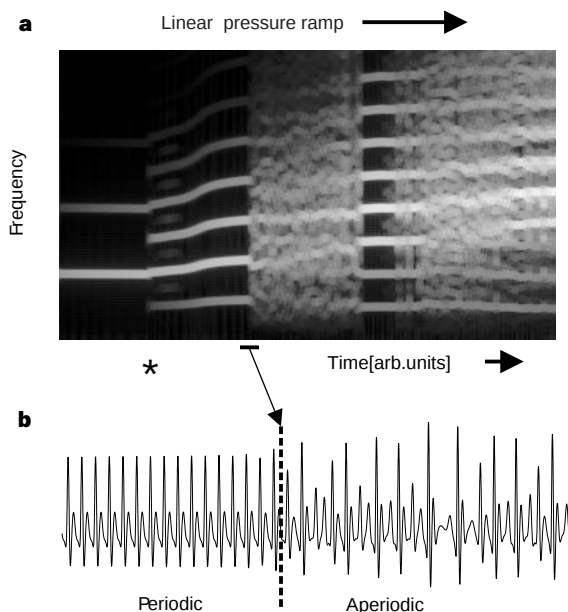


Figure 4 Results of a simple numerical model of airflow through a compliant constriction. The syringeal membranes are modelled as a system of coupled masses and springs. **a**, Spectrogram of the oscillatory membrane displacement. The parameter associated with bronchial pressure is ramped slowly and linearly up from zero. Note the period doubling events (asterisk), and the rapid transition to 'chaotic' dynamics. **b**, The membrane displacement shown expanded around the latter transition showing that the transition occurs rapidly, that is, in approximately one oscillatory period.

tubes¹⁸. The model is formed by a set of ordinary differential equations describing the mechanical motion of the syringeal membranes in the presence of airflow through the bronchial tube. The essential nonlinearity in the model is the nonlinear pressure-flow relation associated with the Bernoulli force. The membranes were modelled as a two-mass system. Membrane parameters were based on estimates of the resonant mode structure of the LI and the MTM (M.S.F. and P.P.M., unpublished results). The essential result is that the model exhibits qualitatively similar behaviour to *in vivo* zebra-finch song as well as the *in vitro* syringeal oscillations. The model shows period doubling and transitions from periodic to chaotic dynamics (Fig. 4), as well as mode-locking transitions (not shown). Our model results suggest a mechanism for the observed mode locking, namely coupling of the basic Bernoulli oscillation to a higher vibrational mode in the membranes. An additional mechanism that might underlie mode locking is the coupling of the oscillatory mechanism to tracheal resonances.

The generic signatures of nonlinear dynamics described here have also been described in models of human vocalizations¹⁹, and in patients with vocal disorders²⁰. A distinction with human vocalizations demonstrated by our model and the *in vitro* results is that detailed spectral shaping in zebra-finch song arises in large part from syringeal dynamics¹, as opposed to filtering effects of the vocal tract.

We have demonstrated that the isolated syrinx behaves as a classic nonlinear dynamical system, often exhibiting rapid transitions between distinct oscillatory states. The nonlinear behaviour of the zebra-finch syrinx results from the large amplitude motion of the vibrating membranes and the involvement of the relatively heavy LI. As a consequence, simple trajectories in a small number of control parameters can result in a complex sequence of acoustic signals. Thus, in addition to the central song control system, the syrinx must also be viewed as an important component in the hierarchy of structures determining the temporal organization of birdsong. Furthermore, the resulting complexity of the mapping from articu-

lators to acoustic output suggests that song learning may require a sophisticated internal model of syringeal dynamics. □

Methods

Isolated syrinx. Adult male zebra finches (Canary Bird Farm, Old Bridge, New Jersey) were anaesthetized and killed by decapitation. The syrinx was removed, cleaned of excess connective tissue, and mounted in the experimental chamber. The trachea was connected to an adjustable vacuum source, and one bronchial tube was connected to a source of humid air with adjustable pressure. The other bronchial tube was pinched closed and sealed with cyanoacrylate. The trachea and anterior portion of the syrinx was embedded in 3% agarose in 10 mM phosphate-buffered saline. Tracheal and bronchial pressures were recorded with electronic pressure sensors (Sensym). Above a threshold flow rate (~ 0.5 standard litres per minute), the LI and the MTM oscillated spontaneously over a wide range of flow rates (0.5–2.5 standard litres per minute) and pressure differentials between the trachea and bronchus (0.1–2.0 kPa). Oscillation at pressures or flow rates higher than this physiological range resulted in rupture of the medial membranes. Acoustic signals generated by the syrinx were recorded using either an electret microphone placed ~ 2 cm from the syrinx, or with a probe microphone (Etymotic Research) placed into the trachea.

Syrinx response to parameter changes. Dependence of the oscillatory state of the excised syrinx on parameters was examined for variations in both tracheal pressure and membrane tension. To study pressure variations, the bronchial inlet was held at a constant pressure (0–0.5 kPa) while the tracheal outlet pressure was smoothly ramped down 1–2 kPa below atmospheric pressure, and then ramped back to atmospheric pressure. The control pressures were measured ~ 50 cm from the syrinx to reduce the sensitivity of the pressure measurement to changes in syringeal resistance. The dependence of oscillatory state on the position of the LE was examined during sustained syringeal oscillation at a constant tracheal pressure. By using forceps, the LE was forced manually into the bronchial lumen until the oscillation ceased. Simultaneous video recording of the preparation and the acoustic signal enabled later quantification of LE displacement and the associated acoustic modulations. Displacement of the LE produced no apparent displacement or tensioning of the LI or MTM, so the frequency changes were probably a result purely of constriction of the bronchial lumen, rather than tension changes.

Stroboscopic imaging. Stroboscopic images of the oscillating syrinx were obtained *in vitro* to examine in detail the motion of the LI and the MTM. The syrinx was induced to oscillate as described above. Pulsed illumination (10 μ s duration) from a cluster of four bright light-emitting diodes (LEDs) was synchronized with the fundamental oscillation period of the membrane as follows. The measured acoustic signal was rectified, band-pass filtered around the oscillation frequency (typically 500–1,000 Hz), and used to trigger a variable-delay pulse generator. The pulse generator triggered the LEDs at a known time relative to the acoustic signal. The delay was slowly swept through several milliseconds to sample all phases of the oscillation; the images were recorded with a CCD camera and video-recorded with a Sony Betacam recorder for later analysis.

Spectral analysis. Spectrograms were computed with direct multitaper spectral estimates²¹. In Fig. 1c, superimposed on the spectrogram is the F-spectrum²¹. This quantity measures the goodness-of-fit to a sinusoid at any given frequency. The F-spectrum was computed with a moving window, and the spectral components exceeding a fixed confidence level were superimposed on the displayed spectrogram.

Received 8 May; accepted 29 June 1998.

- Greenewalt, C. H. *Bird Song: Acoustics and Physiology* (Smithsonian Institution Press, Washington DC, 1968).
- Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary, *Serinus canarius*. *J. Comp. Neurol.* **165**, 457–486 (1976).
- Nottebohm, F., Kelley, D. B. & Paton, J. A. Connections of vocal control nuclei in the canary telencephalon. *J. Comp. Neurol.* **207** (1982).
- McCasland, J. S. Neuronal control of bird song production. *J. Neurosci.* **7**, 23–39 (1987).
- Scharff, C. & Nottebohm, F. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *J. Neurosci.* **11**, 2896–2913 (1991).
- Konishi, M. Pattern generation in birdsong. *Curr. Opin. Neurobiol.* **4**, 827–831 (1994).
- Vu, E., Mazurek, M. E. & Kuo, Y.-C. Identification of a forebrain motor programming network for the learned song of zebra finches. *J. Neurosci.* **14**, 6924–6934 (1994).
- Yu, A. & Margoliash, D. Temporal hierarchical control of singing in birds. *Science* **273**, 1871–1875 (1996).
- Immelman, K. in *Bird Vocalizations* (ed. Hinde, R. A.) pp. 61–74 (Cambridge Univ. Press, 1969).

10. Price, P. H. Developmental determinants of structure in zebra finch song. *J. Comp. Physiol. Psychol.* **93**, 260–277 (1979).
11. Fletcher, N. H. Mode locking in non-linearly excited inharmonic musical oscillators. *J. Acoust. Soc. Am.* **64**, 1566–1569 (1978).
12. Jackson, E. A. *Perspectives of Nonlinear Dynamics* (Cambridge Univ. Press, 1992).
13. Hartley, R. S. & Suthers, R. A. Airflow and pressure during canary song: direct evidence for mini-breaths. *J. Comp. Phys. A* **165**, 15–26 (1989).
14. Goller, F. & Suthers, R. A. Implications for lateralization of bird song from unilateral gating of bilateral motor patterns. *Nature* **373**, 63–66 (1995).
15. Goller, F. & Larsen, O. N. A new mechanism of sound generation in songbirds. *Proc. Natl Acad. Sci. USA* **94**, 14787–14791 (1997).
16. Fletcher, N. H. Bird song—a quantitative acoustic model. *J. Theor. Biol.* **135**, 455–481 (1988).
17. Ishizaka, K. & Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Syst. Tech. J.* **51**, 1233–1268 (1972).
18. Bertram, C. D. & Pedley, T. J. A mathematical model of unsteady collapsible tube behaviour. *J. Biomech.* **15**, 39–50 (1982).
19. Berry, D. A., Herzel, H., Titze, I. R. & Krischer, K. Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *J. Acoust. Soc. Am.* **95**, 3595–3604 (1994).
20. Herzel, H., Berry, D. A., Titze, I. R. & Saleh, M. Analysis of vocal disorders with methods from nonlinear dynamics. *J. Speech Hear. Res.* **37**, 1008–1019 (1994).
21. Thompson, D. J. Spectrum estimation and harmonic analysis. *Proc. IEEE* **70**, 1055–1096 (1982).

Acknowledgements. We thank W. Denk, M. Konishi and S. S.-H. Wang for comments on the manuscript.

Correspondence and requests for materials should be addressed to M.S.F. (e-mail: fee@lucent.com).

When temporal terms belie conceptual order

Thomas F. Münte^{*‡}, Kolja Schiltz^{*} & Marta Kutas^{*†}

^{*}Department of Cognitive Science and [†]Department of Neurosciences, University of California, San Diego, La Jolla, California 92093-0515, USA

[‡]Department of Neurology, Medizinische Hochschule Hannover, 30623 Hannover, Germany

We conceive of time as a sequential order of real-world events, one event following another from past to present to future. This conception colours the way we speak of time (“we look forward to the time”) and, as we show here, the way we process written statements referring to the temporal order of events, in real time. Terms such as ‘before’ and ‘after’ give us the linguistic freedom to express a series of events (real or imaginary) in any

order. However, sentences that present events out of chronological order require additional discourse-level computation. Here we examine how and when these computations are carried out by contrasting brain potentials across two sentence types that differ only in their initial word (‘After’ X, Y versus ‘Before’ X, Y). At sites on the left frontal scalp, the responses to ‘before’ and ‘after’ sentences diverge within 300 ms; the size of this difference increases over the course of the sentences and is correlated with individual working-memory spans. Thus, we show that there are immediate and lasting consequences for neural processing of the discourse implications of a single word on sentence comprehension.

There is no common agreement of how conceptual knowledge in long-term memory (LTM) is used to create a mental representation of a sentence during online comprehension. However, there is a consensus that we create a temporary representation of a sentence and the situation it refers to^{1–3}. In addition, words must activate world knowledge, although when this occurs remains controversial^{1–3}. To examine this, we contrasted event-related brain potentials (ERPs) during the reading of two sentence types whose processing we expected to differ based on the linguistic and conceptual knowledge activated by their initial words (‘before/after’).

Both sentence types consist of an initial subordinate clause and a subsequent main clause, each describing a distinct event that is neither logically nor causally related to the other (for example, ‘Before/After’ the psychologist submitted the article, the journal changed its policy). In other words, the event in each clause could easily be understood without reference to the other. However, we believe that the first word leads readers to expect some non-arbitrary relationship between the two events. This expectation is based on real-world knowledge of time, and linguistic knowledge of how these notions map onto expressions of time. Our experiences in the real world suggest to us that time unfolds sequentially, with current events sometimes causing future events. Our linguistic knowledge tells us that temporal conjunctions often draw attention to the sequence of events in a discourse^{4,5}. Moreover, ‘before’ and ‘after’ access such a sequence from different starting points: ‘after’

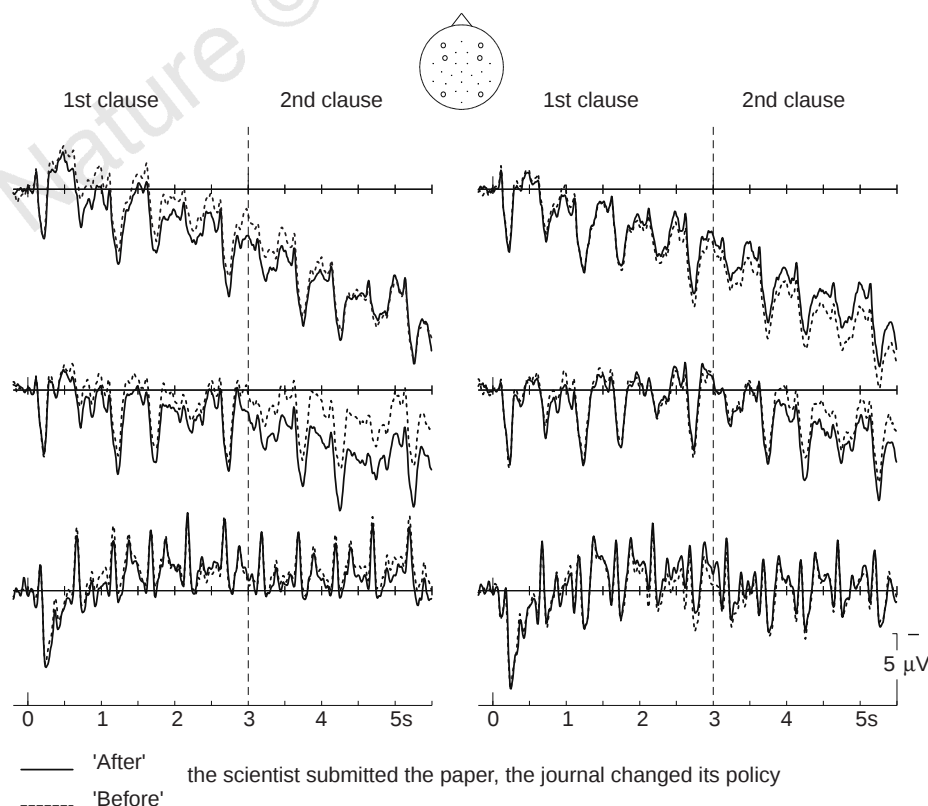


Figure 1 Grand average ($n = 24$) cross-sentence ERPs from prefrontal, frontal and occipital scalp sites (open circles on schematic head) of the left and right hemispheres (left and right sides of the figure, respectively) time-locked to ‘before’ or ‘after’. ‘Before’ sentences are more negative than ‘after’ sentences throughout (main effect of sentence type: clause 1, $F_{1,23} = 9.98$, $P < 0.005$; clause 2, $F_{1,23} = 11.23$, $P < 0.003$), especially over the left compared with the right hemisphere sites (sentence type by site by hemisphere interaction: clause 1, $F_{10,230} = 4.63$, $P < 0.001$; clause 2, $F_{10,230} = 5.65$, $P < 0.001$). The difference between sentence types is reliably larger during the second than the first clause (main effect of clause for ‘before/after’ difference: $F_{1,23} = 4.78$, $P < 0.04$).

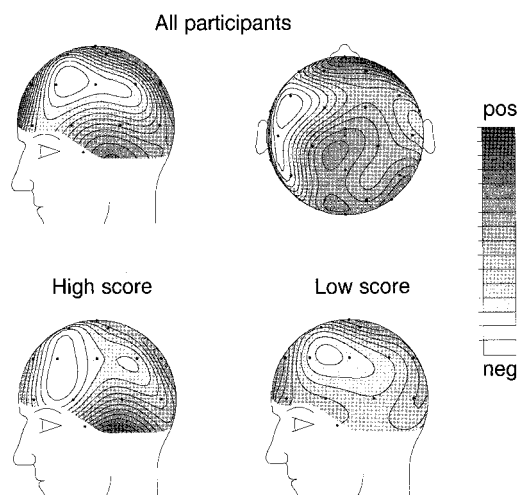


Figure 2 Spline-interpolated isovoltage maps²⁰ displaying the mean difference between 'before' and 'after' sentences (mean amplitude from 500 to 5500 ms). 'Before' sentences lead to a more negative response that is maximal over the left anterior scalp. Although the effect is different in amplitude for individuals with high versus low working-memory scores, its distribution is very similar (group by site interaction after rescaling according to McCarthy and Wood²¹: $F_{21,294} = 0.81$, not significant). Note that a relative scaling is used (min and max: all participants, -1.7 and 1.0 μV ; high score, -3.2 and -0.1 μV ; low score, -0.9 and 2.0 μV).

signals that events will be expressed in their actual order of occurrence, whereas 'before' signals that events will be expressed in reverse order. If real-world knowledge has an influence on sentence processing, then the responses to 'before' and 'after' sentences should diverge even though they differ only minimally in their initial lexical item. We suggest that the discourse representations are structured by real-world knowledge of temporal sequence. 'After' sentences adhere to this default model, whereas 'before' sentences do not; as a consequence, 'before' sentences require additional computations.

Consistent with non-interactive models of sentence comprehension, this processing difference could occur relatively late in the second clause, that is, at the point where all of the information relevant for the temporal comparison from each clause is available for integration in a message-level representation^{1,2}. However, in line with more interactive models of language comprehension^{1,2}, we propose that conceptual knowledge is used continuously in real time to build a discourse-level representation of a sentence. Thus, 'before' and 'after' sentences should diverge from their outset.

As shown in Fig. 1, the responses to 'before' and 'after' sentences do diverge as early as 300 ms into the processing of the first word of the sentence; 'before' sentences elicit greater negativity. This difference in brain potential has a left anterior focus (Fig. 2) and grows progressively larger across the sentence. It thus appears that words do access conceptual knowledge as part of the comprehension process almost immediately and this has lasting processing consequences.

These findings also may relate to the possible locus of these effects in the cognitive system, as a similar ERP response has been observed in contrasts of written and spoken sentence types that differ in working-memory demands⁶⁻⁸. In several experiments, according to

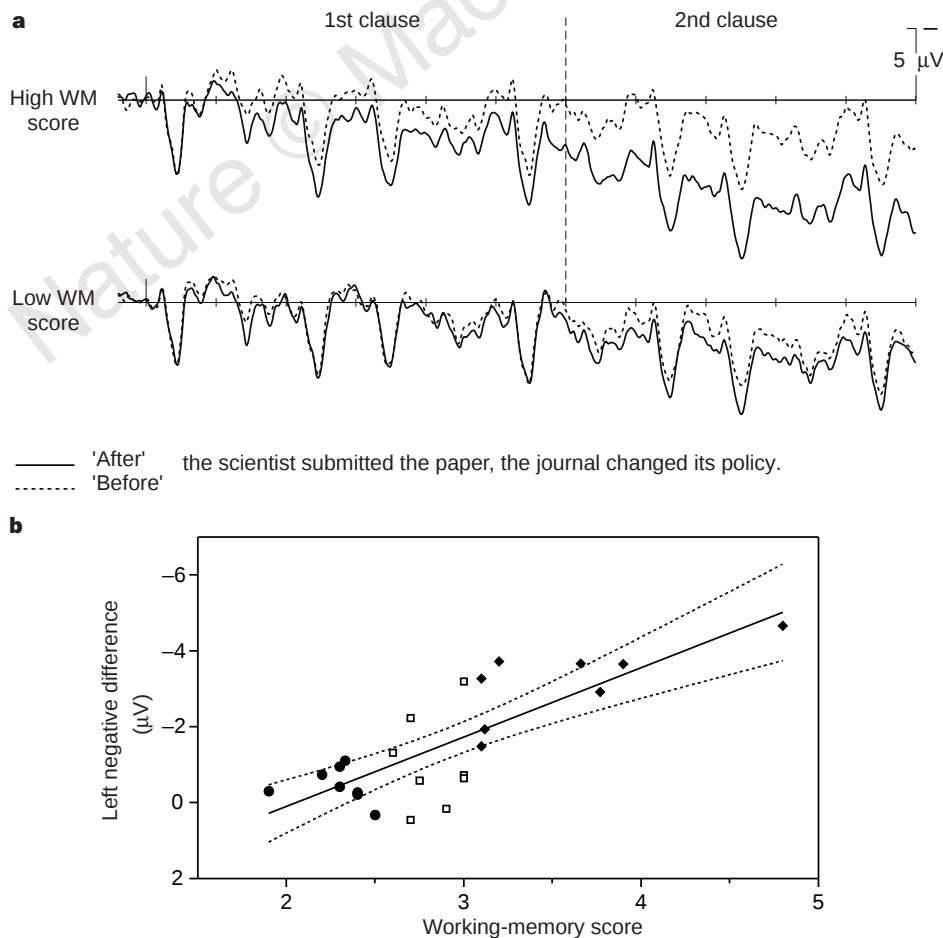


Figure 3 Relationship of left negative difference and working memory span. **a**, Cross-sentence ERPs from the left frontal recording site for the individuals with the highest ($n = 8$; mean = 3.59, s.d. = 0.59) and lowest ($n = 8$; mean = 2.29, s.d. = 0.18) scores of working-memory (WM) span. Individuals with higher working memory span show a more pronounced difference between 'before' and 'after' sentences (group by sentence type interaction: clause 1, $F_{1,14} = 9.61$, $P < 0.008$; clause 2, $F_{1,14} = 8.37$, $P < 0.015$). **b**, The amplitude of the left frontal negativity (mean difference between 'before' and 'after' sentences across the sentence) is significantly correlated with working memory span scores ($r = 0.783$, $F_{1,22} = 35.4$, $P < 0.0001$). This correlation remains significant even after omission of the subject with the best score ($r = 0.727$, $F_{1,21} = 23.6$, $P < 0.0001$). Filled circles, low score; open squares, medium score; filled diamonds, high score.

psycholinguistic analyses, sentences that are more demanding of working memory are associated with greater negativity over left frontal regions than those that are less demanding^{6–8}. These differences have a scalp distribution consistent with the proposed involvement of the frontal regions in working-memory processes and are more pronounced in good than in poor comprehenders^{9–13}. If the 'before/after' difference in our data reflects that same process, it too should reflect individual differences. This was indeed found: the largest differences are observed in individuals with high scores for working-memory span (Fig. 3a). In fact, the size of this slow potential difference for 'before' versus 'after' sentences is highly correlated with working-memory span (Fig. 3b).

In previous studies, working-memory load was manipulated by contrasting sentences with different syntactic structures. In some cases, the sentences also differed in the order of the lexical items^{7,8}. In our study, the critical sentences differ only in the initial word; their syntactic structure and lexical content are otherwise identical. Thus, the prolonged ERP difference must be due to the different information that 'before' and 'after' activate in long-term memory and the computational consequences of these for the creation of a discourse representation, presumably in working memory. It is this difference that we suggest is reflected in the ERP effect (left frontal negativity) we observe. Our comparison is a purer demonstration of the proposed link between this left frontal negativity and working-memory processes. Working memory also has been implicated in the difficulties that children have in processing temporal terms^{14–16}, and in the comprehension deficits that adults with Parkinson's disease experience with 'before' compared with 'after' sentences¹⁷.

In summary, these results show the power of a single word not only to express a concept but also to affect sentence processing in real time. To our knowledge, this is the first demonstration that high-level, real-world knowledge has immediate and sustained consequences on neural processing during sentence comprehension. □

Methods

Stimuli. Each subject read 120 critical (60 'before', 60 'after') and 480 filler sentences presented in random order, one word at a time (duration 200 ms, with 500 ms between word onsets) in yellow in the middle of a video monitor. The critical sentences had the following structure: Before/After the noun1 verb1-ed the noun2, the noun3 verb2-ed the noun4 Word frequency was determined using the CELEX corpus¹⁸ and ranged between 0 and 1474 per million (median = 8.6) for the nouns and between 0 and 2372 per million (median = 27.3) for the verbs, most of which were strongly transitive, with a few having possible ditransitive or intransitive readings. 'Before' and 'after' versions of each sentence were shown to different halves of the 24 right-handed, English-speaking volunteers (11 women; mean age 21.7 years); thus, across subjects the exact same sentences occurred in the 'before' and 'after' conditions. Each critical sentence was followed by a true/false comprehension probe for which the original sentence was rephrased by changing either the temporal conjunction, the order of the clauses, the position of temporal conjunction (sentence initial versus between clauses), or all three. The fillers comprised: 120 sentences containing an embedded object-relative clause; 100 sentences starting with a proper name; 180 two-clause sentences beginning with 'although', 'as' or 'because'; and 80 sentences of varying structures including sentences of one clause.

Assessment of working memory span. Scores on a test of working memory span¹⁹, which assess individuals' capacity to process and store verbal information simultaneously, were used to group subjects (high, medium and low). For this test, subjects read increasingly larger (2, 3, 4, 5 and 6) sets of lengthy sentences aloud; immediately after each set, subjects attempt to recall the last word of each sentence in order (maximum score = 6).

Recording and analysis of biosignals. ERPs (time constant = 8 s) were recorded using standard recording procedures from 26 geodesically spaced sites on the scalp^{7,8}. ERPs were obtained for 6144 ms epochs starting 300 ms before onset of the sentence. ERPs were quantified by mean amplitude measures for clause 1 (500–3,000 ms) and clause 2 (3,000–5,500 ms) relative to baseline. Subsequent ANOVA statistics used the Huynh–Feldt correction.

Received 12 February; accepted 11 June 1998.

1. Carpenter, P. A., Miyake, A. & Just, M. A. Language comprehension: sentence and discourse processing. *Annu. Rev. Psychol.* **46**, 91–120 (1995).
2. Graesser, A. C., Millis, K. K. & Zwaan, R. A. Discourse comprehension. *Annu. Rev. Psychol.* **48**, 163–189 (1997).
3. McKoon, G. & Ratcliff, R. Memory-based language processing: psycholinguistic research in the 1990s. *Annu. Rev. Psychol.* **49**, 25–42 (1998).
4. Turner, M. *Reading Minds: The Study of English in the Age of Cognitive Science* (Princeton Univ. Press, 1991).
5. Langacker, R. W. *Foundations of Cognitive Grammar* (Stanford Univ. Press, 1987).
6. Mecklinger, A., Schriefers, H., Steinhauer, K. & Friederici, A. D. Processing relative clauses varying on syntactic and semantic dimensions: an analysis with event-related potentials. *Memory Cogn.* **23**, 477–494 (1995).
7. King, J. W. & Kutas, M. Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *J. Cogn. Neurosci.* **7**, 376–395 (1995).
8. Müller, H. M., King, J. W. & Kutas, M. Event-related potentials elicited by spoken relative clauses. *Cogn. Brain Res.* **5**, 193–203 (1997).
9. Goldman-Rakic, P. S. The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. *Phil. Trans. R. Soc. Lond. B* **351**, 1445–1453 (1996).
10. Owen, A. M. The functional organization of working memory processes within human lateral frontal cortex: the contribution of functional neuroimaging. *Eur. J. Neurosci.* **9**, 1329–1339 (1997).
11. Petrides, M. Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Phil. Trans. R. Soc. Lond. B* **351**, 1455–1461 (1996).
12. Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F. & Thulborn, K. R. Brain activation modulated by sentence comprehension. *Science* **274**, 114–116 (1996).
13. Stromswold, K., Caplan, D., Alpert, N. & Rauch, S. Localization of syntactic comprehension by positron emission tomography. *Brain Lang.* **52**, 452–473 (1996).
14. Macaruso, P., Bar-Shalom, E., Crain, S. & Shankweiler, D. Comprehension of temporal terms by good and poor readers. *Lang. Speech* **32**, 45–67 (1989).
15. Natsopoulos, D. & Abadzi, H. Understanding linguistic time sequence and simultaneity: A literature review and some new data. *J. Psycholing. Res.* **15**, 243–273 (1986).
16. Trosborg, A. Children's comprehension of "before" and "after" reinvestigated. *J. Child Lang.* **9**, 381–402 (1982).
17. Natsopoulos, D., Mentenopoulos, G., Bostantzopoulou, S., Katsarou, Z., Grouios, G. & Logothetis, J. Understanding of relational time terms before and after in parkinsonian patients. *Brain Lang.* **40**, 444–458 (1991).
18. Baayen, R. H., Piepenbrock, R. & van Rijn, H. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, Univ. Pennsylvania (1993).
19. Daneman, M. & Carpenter, P. A. Individual differences in working memory and reading. *J. Verb. Learn. Verb. Behav.* **19**, 450–466 (1980).
20. Perrin, F., Pernier, J., Bertrand, O., Giard, M. H. & Echallier, J. F. Mapping of scalp potentials by surface spline interpolation. *Electroencephalogr. Clin. Neurophysiol.* **66**, 75–81 (1987).
21. McCarthy, G. & Wood, C. C. Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroencephalogr. Clin. Neurophysiol.* **62**, 203–208 (1985).

Acknowledgements. We thank K. Federmeier, R. Schul, B. Schmitt, J. King and R. Kluender for their comments. This research was supported by the NICHD, NIA, and the Hermann and Lilly Schilling Foundation, Essen, Germany.

Correspondence and requests for materials should be addressed to T.F.M. (e-mail: muenste.thomas@mh-hannover.de).

Spontaneous pinwheel annihilation during visual development

F. Wolf & T. Geisel

Max-Planck-Institut für Strömungsforschung, Postfach 2853, D-37018 Göttingen and SFB Nichtlineare Dynamik, Universität Frankfurt, Germany

Neurons in the visual cortex respond preferentially to edge-like stimuli of a particular orientation¹. It is a long-standing hypothesis that orientation selectivity arises during development through the activity-dependent refinement of cortical circuitry^{2–4}. Unambiguous evidence for such a process has, however, remained elusive^{5–7}. Here we argue that, if orientation preferences arise through activity-dependent refinement of initially unselective patterns of synaptic connections, this process should leave distinct signatures in the emerging spatial pattern of preferred orientations. Preferred orientations typically change smoothly and progressively across the cortex¹. This smooth progression is disrupted at the centres of so-called pinwheels^{8,9}, where neurons exhibiting the whole range of orientation preferences are located in close vicinity¹⁰. Assuming that orientation selectivity develops through a set of rules that we do not specify, we demonstrate mathematically that the spatial density of pinwheels is rigidly constrained by basic symmetry principles. In particular, the

Box 1 Theory of pinwheel formation.

Random patterns of orientation columns. An ensemble of patterns of orientation columns $\mathbf{z}(\mathbf{x})$ is characterized by the correlation function $C(\mathbf{x}, \mathbf{y}) = \langle \mathbf{z}(\mathbf{x}) \mathbf{z}(\mathbf{y}) \rangle$, where $\langle \rangle$ denotes the average over the ensemble and $\langle \mathbf{z}(\mathbf{x}) \rangle = 0$. If $C(\mathbf{x}, \mathbf{y})$ depends only on the distance between columns, $C(\mathbf{x}, \mathbf{y}) = C(|\mathbf{x} - \mathbf{y}|) = C(r)$, and no additional prior knowledge is assumed, the maximum entropy ensemble of patterns with isotropic power spectrum $P(k) = \int d^2\mathbf{x} C(\mathbf{x}) e^{i\mathbf{k} \cdot \mathbf{x}}$ must be considered. This is an ensemble of gaussian random fields.

The typical spacing of IODs, Λ , in a repetitive pattern can be defined by the mean wavenumber through $\Lambda = 2\pi/\bar{k}$, where $\bar{k} = \int_0^\infty dk k P(k)$ and $P(k)$ is normalized such that $\int_0^\infty dk P(k) = 1$. Pinwheel centres are the zero values of $\mathbf{z}(\mathbf{x})$. Their spatial density is given by $\rho = (1/4\pi) (\int_0^\infty dk k^2 P(k) / \int_0^\infty dk k P(k))$ (ref. 19). As

$$\rho = \frac{1}{4\pi} \frac{\int_0^\infty dk k^3 P(k)}{\int_0^\infty dk k P(k)} = \frac{\bar{k}^2}{4\pi} \frac{\int_0^\infty dk k^3 P(k)}{\left(\int_0^\infty dk k P(k)\right)^2}$$

and $\int_0^\infty dk k^3 P(k) \geq (\int_0^\infty dk k P(k))^2$ (Jensen's inequality²⁸), this density is bounded from below and $\rho = (\pi/\Lambda^2)(1 + \alpha)$, with $\alpha \geq 0$ for every $P(k)$; α vanishes if, and only if, $P(k) = \delta(k - \bar{k})$.

Development dynamics. To model the activity-dependent development of orientation preference one must specify in which way a sequence of neural activity patterns, $\mathcal{A}_i$, modifies the orientation preference and selectivity of cortical neurons. This can be idealized by a modification rule

for the field $\mathbf{z}(\mathbf{x})$, $\mathbf{z}_i(\mathbf{x}) \xrightarrow{\mathcal{A}_i} \mathbf{z}_{i+1}(\mathbf{x})$. The change $\delta \mathbf{z}(\mathbf{x}) = \mathbf{z}_{i+1}(\mathbf{x}) - \mathbf{z}_i(\mathbf{x})$ in general depends on the present pattern and on $\mathcal{A}_i$. For simplicity, we assume here that $\mathbf{z}(\mathbf{x})$ and $\mathcal{A}$ are sufficient to determine $\delta \mathbf{z}(\mathbf{x}) = f(\mathbf{z}(\mathbf{x}), \mathcal{A})$. If $\max|\delta \mathbf{z}(\mathbf{x})|$ is much smaller than the final amplitude of the pattern, $\mathbf{z}(\mathbf{x})$, and the sequence of activity patterns, $\mathcal{A}_i$, are randomly distributed according to a probability distribution, then the change of $\mathbf{z}(\mathbf{x})$ that results from a large number of activity patterns can be approximately described by equation (2), where t is time and $F[\mathbf{z}(\mathbf{x})] = \langle f(\mathbf{z}(\mathbf{x}), \mathcal{A}) \rangle_{\mathcal{A}}$ (ref. 29). Specific examples of such models are considered below; see also Supplementary Information.

Symmetries and pinwheel proliferation. Equation (2) exhibits symmetries a-c if

$$F[\tilde{\mathbf{z}}] = \tilde{\mathbf{z}} F[\mathbf{z}] \quad \text{with} \quad \tilde{\mathbf{z}}(\mathbf{x}) = \mathbf{z}(\mathbf{x} + \mathbf{y}) \quad (\text{a})$$

$$F[\tilde{\mathbf{R}}\mathbf{z}] = \tilde{\mathbf{R}} F[\mathbf{z}] \quad \text{with} \quad \tilde{\mathbf{R}}\mathbf{z}(\mathbf{x}) = \mathbf{z}\left(\begin{bmatrix} \cos(\beta) & \sin(\beta) \\ -\sin(\beta) & \cos(\beta) \end{bmatrix} \mathbf{x}\right) \quad (\text{b})$$

and

$$F[e^{i\alpha}\mathbf{z}] = e^{i\alpha} F[\mathbf{z}]. \quad (\text{c})$$

The symmetries a-c determine many properties of equation (2); c implies that the unselective state $\mathbf{z}(\mathbf{x}) = 0$ is a stationary solution: $F[0] = e^{i\alpha} F[0] \Rightarrow F[0] = 0$.

In the vicinity of this state, equation (2) is approximated by

$$\frac{\partial}{\partial t} \mathbf{z}(\mathbf{x}) = \tilde{\mathcal{L}} \mathbf{z}(\mathbf{x}) + \xi(\mathbf{x}, t), \quad (\text{4})$$

where $\tilde{\mathcal{L}}$ is a linear operator. As a consequence of (a, b), $\tilde{\mathcal{L}}$ is diagonal in Fourier representation and its eigenvalues $\lambda(k)$ depend only on the modulus of the wave vector $k = |\mathbf{k}|$. Equation (2) gives rise to a pattern with a typical spacing of IODs if there is only one interval (k_l, k_h) of wavenumbers with positive eigenvalues. In this case, equation (4) describes the emergence of a pattern of IODs from an unselective initial state, $\mathbf{z}_0(\mathbf{x}) \approx 0$. Equation (4) is solved in terms of the Green's function $G(\mathbf{x}, t) = (1/2\pi) \times \int d^2\mathbf{k} e^{-i\mathbf{k} \cdot \mathbf{x} + \lambda(k)t}$ by $\mathbf{z}_L(\mathbf{x}, t) = \int d^2\mathbf{y} \int_0^t dt' G(\mathbf{y} - \mathbf{x}, t - t') (\xi(\mathbf{y}, t') + \mathbf{z}_0(\mathbf{x}) \delta(t'))$. If $\xi(\mathbf{y}, t')$ and $\mathbf{z}_0(\mathbf{x})$ are random fields, $\mathbf{z}_L(\mathbf{x}, t)$ represents the sum of a set of random variables. According to the central limit theorem²⁸, it will therefore realize a gaussian random field under a wide range of conditions (see Supplementary Information). $\square$

spatial density of pinwheels, which emerge when orientation selectivity is first established, is larger than a model-independent minimal value. As a consequence, lower densities, if observed in adult animals, are predicted to develop through the motion and annihilation of pinwheel pairs.

Symmetries, invariances of a system's behaviour under a group of geometric transformations, are central to the mathematical description of processes throughout nature. Symmetries and symmetry breaking may also represent important aspects of visual cortical development^{3,4}. Early in development, most neurons respond rather unselectively to stimuli of all possible orientations^{11,12}. This relative invariance (symmetry) of responses is broken at around the time of eye opening and neurons acquire preferences for particular orientations^{11,12}. This process also breaks the spatial homogeneity (translation symmetry) of the system by establishing a spatial pattern of orientation preferences (see, for example, ref. 13). The

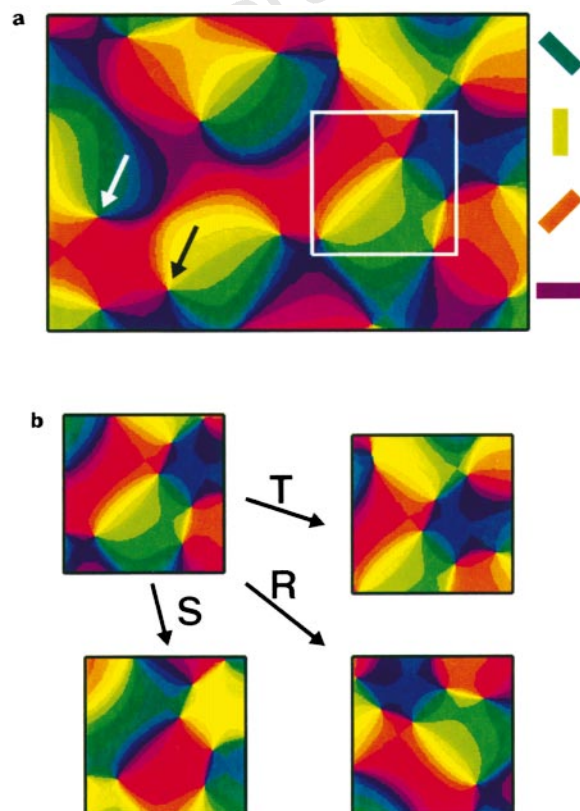


Figure 1 Pinwheels frequently occur in random patterns of IODs. **a**, Numerically synthesized instance of an isotropic and homogeneous gaussian random pattern of IODs. Colours code preferred orientations, as indicated by the bars. Two pinwheel centres are marked by arrows. The white box encloses an area of size Λ^2 , where Λ is the typical spacing of IODs as defined below. The box encloses five pinwheels. The average number of pinwheels enclosed in an area of this size is in general larger than π . The pattern was synthesized by the superposition of 256 plane waves of wavelength Λ distributed homogeneously over spatial directions with coefficients $A(\mathbf{k})$, chosen as complex gaussian random numbers $\mathbf{z}_{RW}(\mathbf{x}) = \sum_{\mathbf{k}=(2\pi/\Lambda)} A(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}}$. Random patterns synthesized in this manner are qualitatively similar to experimentally observed patterns (see ref. 30, for example). **b**, Symmetries of isotropic and homogeneous random patterns. The top left panel shows a pattern of IODs; the three panels marked by arrows show patterns derived by translation (T), by rotation (R), or by shifting all orientation preferences by the same angle (S). If chosen randomly from a homogeneous and isotropic ensemble, all four patterns will emerge with the same probability. Conversely, if all patterns that are related by translations, rotations and orientation shifts emerge with the same probability, the ensemble of patterns must be statistically homogeneous and isotropic.

theoretical significance of symmetry principles derives from the fact that those aspects of a system's behaviour that result from basic symmetries are insensitive to the exact nature of its components' interactions. For instance, in a turbulent flow the distribution of energy among eddies of different sizes is highly constrained by the basic symmetries of the Navier–Stokes equation and is largely insensitive to the detailed structure of the flow¹⁴. In neurobiology, theoretical predictions that are insensitive to the exact nature of neuronal interactions are highly desirable because many interactions either remain unknown or have been characterized only qualitatively, and may differ among species or even individuals. Indeed, the detailed mechanisms that underlie orientation selectivity *per se*, as well as those that guide the refinement of synaptic connections *in vivo*, are still controversial^{15–17}. We now present a symmetry-based theoretical analysis of the dynamics of pinwheels during visual development. First, we shall develop a probabilistic picture of the emergence of patterns of orientation preferences. Within this picture, symmetry assumptions imply that a minimal density of pinwheels must emerge when orientation selectivity is first established. We shall then show mathematically that this treatment accurately represents the dynamics of a large class of models for the activity-dependent development of orientation selectivity. The dynamics of pinwheels in this class of models is therefore highly constrained by symmetry principles and predicts distinct, robust and experimentally verifiable signatures of an activity-dependent generation of orientation preferences.

The spatial pattern of orientation preferences in visual cortex is essentially two-dimensional. Neurons in columns extending perpendicularly to the cortical surface prefer similar orientations⁵. Parallel to the cortical surface, columns exhibiting the same orientation preference form so-called iso-orientation domains (IODs)⁹ and neurons in neighbouring IODs typically prefer progressively changing orientations^{1,18}. The two-dimensional arrangement of orientation columns parallel to the cortical surface can be represented by a field of complex numbers $z(\mathbf{x}) = |z(\mathbf{x})|e^{2i\theta(\mathbf{x})}$ (ref. 4). Here $\mathbf{x}$ denotes the location of a column parallel to the cortical surface, $\theta(\mathbf{x})$ its preferred orientation, and $|z(\mathbf{x})|$ measures the orientation selectivity of the average response of neurons within the column. Because different neurons in the centre of a pinwheel exhibit the whole range of possible orientation preferences¹⁰, their average response is unselective and $|z(\mathbf{x})|$ vanishes at these locations. Every pinwheel centre is therefore a zero of $z(\mathbf{x})$.

We used this representation to estimate the expected density of pinwheels within a probabilistic framework. Our estimate is based on the assumption that, if orientation preference emerges by the activity-dependent refinement of initially crude patterns of synaptic connections, then random influences—like a random setup of the

initial pattern of connections, or an unpredictable sequence of activity patterns that guides the refinement of synaptic connections—determine the emerging pattern of orientation preferences. Mathematically, this assumption is equivalent to considering the emerging pattern of orientation columns as a realization drawn at random from an ensemble of possible patterns. The formation of a pinwheel is then a joint event. If a given column develops at random a preference, for example for horizontally oriented stimuli, this will happen with a defined probability. There is also a joint probability that neighbouring columns develop preferences for vertical and the two oblique orientations at the same time. The overall result of this event is the formation of a pinwheel. How frequently such a configuration is expected to arise in a given area determines the spatial density of pinwheels in the emerging pattern. To calculate the density, spatial correlations in the emerging pattern of IODs must be taken into account. The emerging orientation preferences of nearby columns must be highly correlated because orientation preferences change continuously across the cortex as soon as the pattern can be visualized experimentally (at about the time of eye opening)¹³. Furthermore, separate IODs of the same orientation preference exhibit a typical spacing, Λ , early on¹³. Considered statistically, this implies that the orientation preferences that emerge in columns, which are one Λ apart, must be positively correlated. To calculate the expected density of pinwheels, ρ , we assumed that these correlations—whatever their exact form—depend only on the distance between columns and completely specify the ensemble. The emerging pattern of IODs then realizes a homogeneous and isotropic gaussian random field (Fig. 1) and the expected density of pinwheels is given by

$$\rho = \frac{\pi}{\Lambda^2} (1 + \alpha) \quad (1)$$

where α is a number that describes the structure of spatial correlations (Box 1): α vanishes if patterns exhibit only a vanishingly small range of possible spacings between IODs and is positive if a finite range of spacings is present. Equation (1) therefore implies that one expects to find on average at least $\pi \approx 3.1$ pinwheels in an area of size Λ^2 .

We used two assumptions to derive equation (1). (1) Two-point correlations in the emerging pattern depend only on the distance between columns; and (2) gaussian statistics: these correlations are sufficient to specify the ensemble of possible early patterns. Assumption (1) is related to a set of symmetries. It is correct if, given that one particular pattern of IODs can emerge, the patterns that result from (a) shifting the pattern parallel to the cortical surface, or (b) rotating the pattern parallel to the cortical surface, or (c) shifting all orientation preferences by the same angle, may in principle form as well, and will arise with the same probability (Fig. 1b). Assumption (2) is likely to be fulfilled if a large number of random factors determine the pattern of IODs (central limit theorem). Qualitatively, this appears very plausible: the elementary event in any model of activity-dependent development is the modification of neuronal selectivities by a pattern of neuronal activity. In the visual cortex, IODs arise over a period of many hours or a few days^{12,13}. The activity patterns that drive this process are presumably only correlated over time intervals of the order of seconds. Therefore the pattern of IODs must arise through the cumulative effect of a large number of independent activity patterns which individually might induce only minor changes.

Any model of visual cortical development that obeys assumptions (1) and (2) will produce a density of pinwheels in accord with equation (1). In particular, the very general class of models defined by a dynamics

$$\frac{\partial}{\partial t} z(\mathbf{x}) = F[z(\cdot)] + \xi \quad (2)$$

and the group of symmetries *a–c* (see Box 1), if beginning from random initial conditions, obeys these assumptions during the

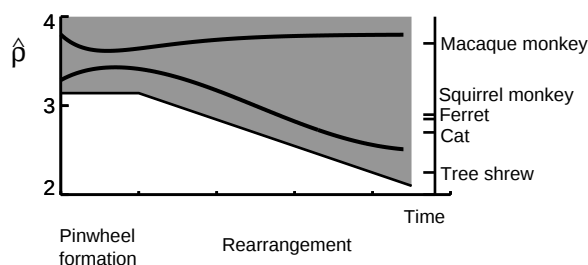


Figure 2 Activity-dependent mechanisms constrain the scaled density of pinwheels during development. Shaded area: accessible range for allowed trajectories of $\hat{\rho}(t)$ as a function of time (*x*-axis) during development according to equations (1, 2). Two possible trajectories are displayed (lines). Right axis: observed scaled pinwheel densities in adult animals of different species (from Table 1). Low densities as observed for example in cats and squirrel monkeys, and tree shrews can only develop through an initial overproduction and subsequent annihilation of pinwheels.

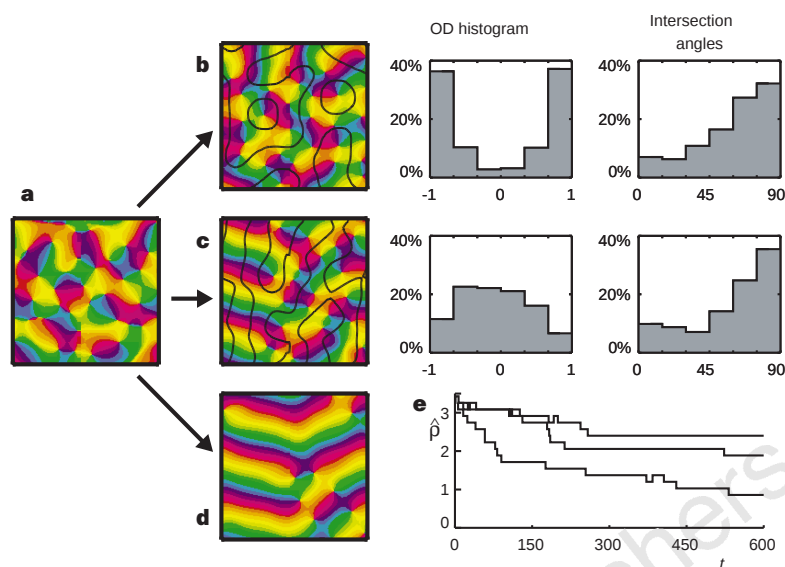


Figure 3 Rearrangement of IODs in the presence or absence of ocular dominance columns. **a**, Early and **b–d**, late patterns of IODs and ocular dominance borders (black lines) in three simulations of the elastic network^{20,21} (see Methods) started from the same initial pattern (in **a**, $t = 10$) but with different degrees of ocular dominance segregation: **b**, strong ocular dominance segregation (most units are monocular); **c**, intermediate ocular dominance segregation (most units are binocular), and **d**, no ocular dominance segregation (all at $t = 600$). Colour coding is as in Fig. 1. The graphs to the right of **b** and **c** show the ocular dominance histogram (left) and the histogram of intersection angles between ocular dominance borders and iso-orientation domains (right) for the patterns in **b** and **c**, respectively. Ocular dominance borders impose a geometric constraint on the pattern of IODs. In both patterns **b** and **c**, iso-orientation domains show a tendency to intersect ocular dominance borders at steep angles: steep intersec-

tion angles ($\psi \approx 90$ deg) are more frequent than shallow angles ($\psi \approx 0$ deg). The simulations show that this constraint impairs pinwheel annihilation more strongly with strong ocular dominance segregation. The ocular dominance histograms show the frequencies of different ocular dominance values o . Ocular dominance values o are binned into 6 equidistant classes. Units that prefer one eye ($o \approx \pm 1$) are most frequent in **b**. Units that are binocular ($o \approx 0$) are most frequent in **c**. **e**, the scaled density of preserved pinwheels $\hat{\rho}$ increases with the degree of ocular dominance segregation. $\hat{\rho}(t)$ is for the three simulations: top, strong ocular dominance segregation; middle, intermediate ocular dominance segregation; bottom, no ocular dominance columns. The x-axis shows the time in units τ (see Methods); orientation selectivity initially arises over a period of a few τ . Steps in **e** represent individual annihilation events.

initial organization of an orientation pattern (Box 1). Therefore, any such model must produce initial patterns with a pinwheel density of $\rho \geq \pi/\Lambda^2$. In equation (2), $\frac{\partial}{\partial t} z(\mathbf{x})$ is the temporal change of $z(\mathbf{x})$ on a slow timescale, $F[\cdot]$ is a nonlinear operator and $\xi(\mathbf{x}, t)$ is a spatiotemporal random process modelling intrinsic and stimulus-induced fluctuations. Here $F[\cdot]$ represents the average of many activity-induced modifications of the pattern of IODs (Box 1). How an individual pattern of IODs transforms according to equation (2) depends on the precise form of the operator $F[\cdot]$ and therefore on the specific rules governing activity-induced changes. However, the kind of patterns that emerge when orientation selectivity is first established from a random unselective initial state is common to all models that exhibit the symmetries *a–c* (Box 1).

It is not implied that lower densities of pinwheels cannot develop within this class of models. However, densities $\rho < \pi/\Lambda^2$ are predicted to form from an early high-density state through a secondary reduction of the number of pinwheels. The only way to reduce the number of pinwheels by continuously rearranging the pattern of IODs is the collision and annihilation of pairs of pinwheels with opposite chirality¹⁹. The considered class of models therefore also predicts that pinwheel densities $\rho < \pi/\Lambda^2$, if found, have been formed through the motion and annihilation of pinwheels (Fig. 2).

Low pinwheel densities have been observed in adult animals. Table 1 summarizes the data available at present on pinwheel densities and the typical spacings of IODs in different species. We used these quantities to estimate the scaled pinwheel density $\hat{\rho} = \rho\Lambda^2$ in each species. The abundance of pinwheels clearly varies across species. Macaque monkeys exhibit the largest and tree shrews the smallest scaled pinwheel densities, whereas squirrel monkeys, ferrets and cats show intermediate values. According to our theory, scaled densities $\hat{\rho} < \pi \approx 3.1$ imply pinwheel annihilation during development (Fig. 2). Pinwheel anni-

hilation is therefore predicted to occur in cat and tree shrew striate cortex. Pinwheel annihilation may also occur in ferrets and squirrel monkeys; it may not occur in macaque monkeys. Furthermore, the data suggest a relation between ocular dominance segregation and pinwheel density. Species in which pronounced ocular dominance columns are present (such as macaque monkeys) exhibit a higher scaled density compared to species in which ocular dominance columns are weak or absent (that is, tree shrews).

As pinwheel annihilation is predicted to occur in several species, the question arises of whether pinwheels do typically move in concrete models of visual development. We will therefore conclude with results from simulations of one such model²⁰. It is assumed in this model that the orientation preferences of cortical units change in response to afferent stimuli, which are described by their location and orientation in visual field coordinates. Cortical units are activated if their receptive-field position and orientation preference are close to the stimulus. Units unselective for stimulus orientation are activated if a stimulus is close to their receptive-field position.



Figure 4 Annihilation of a pair of pinwheels in a simulation of the elastic network (see Methods). Pinwheel annihilation replaces a pair of pinwheels by a linear zone, in which iso-orientation domains form a system of parallel strips. Pinwheels in an annihilating pair are of opposite chirality. In one pinwheel, orientation preference increases clockwise; in the other pinwheel it increases anticlockwise. Colour code as in Fig. 1.

Table 1 Pinwheel density, spacing of iso-orientation domains and OD segregation for different animals

Species	ρ	Λ	$\hat{\rho}$	OD columns
Macaque monkey	8.1	0.68	3.75	Strong
Squirrel monkey	11.1	0.51	2.9	Intermediate
Ferret	5.5	0.75	2.85	Intermediate
Cat	2.1–2.4	1.1	2.5–2.9	Intermediate
Tree shrew	7.9–9.5	0.52	2.1–2.6	Absent

Pinwheel density ρ [1/mm²], typical spacing of iso-orientation domains Λ [mm], scaled density $\hat{\rho} = \rho\Lambda^2$ and degree of ocular dominance (OD) segregation extracted from published data on the primary visual cortex of several species. $\hat{\rho}$ measures the relative abundance of pinwheels by the average number of pinwheels in a region of area Λ^2 . Macaque and squirrel monkey, data from ref. 22; ferret, data from ref. 23; cat, spacing from refs 23, 24, pinwheel density from refs 23, 25. In the tree shrew, extensive regions exhibiting a more stripe-like pattern of iso-orientation domains and consequently low densities of orientation centres have been reported²⁶. One typical orientation map²⁶ exhibits pinwheel densities of 7.9 mm⁻², 8.6 mm⁻² and 9.5 mm⁻² in the caudal, central and rostral part of area 17, respectively, and a typical spacing of 0.52 mm. For comparative reviews of ocular dominance segregation, see refs 18, 27.

The selectivities of activated units are then changed better to match the stimulus. At the same time, neighbouring units interact so that the smoothness of selectivities across the cortical layer is enhanced (see Methods). When starting from an initial condition in which all cortical units are not orientation-selective, application of the same rules can induce the formation of a pattern of IODs. By assigning cortical units an ocular dominance index and introducing stimuli that are either binocular or dominated by one eye, this model has been generalized to describe the coordinated development of orientation and ocular dominance columns²¹. In this case, stimuli are assumed to activate more strongly those units whose ocular dominance matches the stimulus' dominant eye. Analogous to the treatment of orientation preferences, the ocular dominance of activated neurons is then changed towards the eye that dominates the stimulus (see Methods).

In agreement with the theory developed here, more than π/Λ^2 pinwheels proliferate as a pattern of IODs arises from an initially unselective state (Fig. 3a, e). This early pattern is not stable, but rearranges under the influence of continuing stimulus-driven changes. The predominant process during this rearrangement is the motion, collision and annihilation of pairs of pinwheels (Fig. 4). In Fig. 3, three simulations starting from the same early orientation map are compared. These simulations use sets of stimuli that lead to strong (Fig. 3b), intermediate (Fig. 3c), or no ocular dominance segregation (Fig. 3d). The speed and degree of pinwheel annihilation in these simulations reflects the degree of ocular dominance segregation (Fig. 3e). The number of annihilating pinwheel pairs is largest without ocular dominance segregation (Fig. 3d), smaller with an intermediate degree of ocular dominance segregation (Fig. 3c), and least with strong ocular dominance segregation (Fig. 3b). In a large number of similar simulations, we observed that the tendency of pinwheels to annihilate after the initial emergence of orientation selectivity and the ability of ocular dominance segregation to slow down or stop pinwheel annihilation, were independent of model parameters and details (see Supplementary Information). A simple explanation of the observed inter-species differences in the scaled pinwheel density (Table 1) is therefore that species in which ocular dominance columns are weak or absent perform extensive pinwheel annihilation, whereas strong ocular dominance segregation prevents a reduction of the pinwheel density during development.

The development of cortical circuitry is a complex process that involves a multitude of interactions at the molecular, cellular and network level¹⁶. In contrast to this microscopic complexity, our results indicate that an activity-dependent origin of orientation preference has robust and distinct signatures in the dynamics of pinwheels during development. Verification of the prediction that pinwheels move and annihilate in species that exhibit low densities of pinwheels in the adult would be strong evidence for an activity-dependent generation of orientation preferences in the visual cortex. □

Methods

In the elastic network^{20,21}, the field $z(\mathbf{x})$ is complemented by two further fields, $o(\mathbf{x})$ and $r(\mathbf{x})$, describing the pattern of ocular dominances and receptive-field centre positions. $r(\mathbf{x})$ represents the receptive field centre coordinates of a unit at $\mathbf{x}$ by a complex number. Its ocular dominance is described by a real number, $o(\mathbf{x})$, such that $o(\mathbf{x}) > 0$; $o(\mathbf{x}) < 0$ denotes preference for the left (right) eye. These selectivities are modified in response to stimuli described by a set of parameters: $s_z = |s_z|e^{i\theta}$, s_o , $s_r = r_x + ir_y$, where θ denotes the orientation of a stimulus, s_r its position, and s_o describes whether activated units are forced to prefer the left ($s_o > 0$) or right ($s_o < 0$) eye. In response to a stimulus, $z(\mathbf{x})$, $r(\mathbf{x})$ and $o(\mathbf{x})$ were changed according to

$$\delta z(\mathbf{x}) = \epsilon([s_z - z(\mathbf{x})]e_o(\mathbf{x}, \mathbf{S}, z(\cdot), r(\cdot), o(\cdot)) + \eta \Delta z(\mathbf{x}))$$

$$\delta o(\mathbf{x}) = \epsilon([s_o - o(\mathbf{x})]e_o(\mathbf{x}, \mathbf{S}, z(\cdot), r(\cdot), o(\cdot)) + \eta \Delta o(\mathbf{x})) \quad (3)$$

$$\delta r(\mathbf{x}) = \epsilon([s_r - r(\mathbf{x})]e_o(\mathbf{x}, \mathbf{S}, z(\cdot), r(\cdot), o(\cdot)) + \eta \Delta r(\mathbf{x}))$$

where $e_o(\mathbf{x}, \dots)$ represents the cortical activity pattern. It is given by

$$e_o(\mathbf{x}, \mathbf{S}, z(\cdot), r(\cdot), o(\cdot)) = \frac{e^{-(|s_r - r(\mathbf{x})|^2 + |s_z - z(\mathbf{x})|^2 + |s_o - o(\mathbf{x})|^2)/2\sigma^2}}{\int d^2y e^{-(|s_r - r(\mathbf{y})|^2 + |s_z - z(\mathbf{y})|^2 + |s_o - o(\mathbf{y})|^2)/2\sigma^2}}$$

where σ is the receptive field size within the stimulus parameter space. η measures the strength of lateral interactions, Δ is the two-dimensional Laplacian and ϵ defines the amplitude of modifications. The stimulus parameters s_z and s_r were distributed uniformly in the domains $|s_r| < 2$ and $0 \leq \text{Re}(s_r)$, $\text{Im}(s_r) < 8.7$, s_o was distributed either according to a normal distribution $p_n(s_o) = (1/\sqrt{2\pi})e^{-s_o^2/2}$ or a bimodal distribution $p_b(s_o) = \frac{1}{2}\delta(s_o + 1) + \frac{1}{2}\delta(s_o - 1)$. $p_b(s_o)$ typically leads to more pronounced ocular dominance segregation compared to $p_n(s_o)$. For simplicity, the stimulus distributions were chosen such that the characteristic spacing Λ_{OD} of neighbouring ocular dominance columns and Λ_{OP} of IODs are identical (see Supplementary Information). The ability of ocular dominance segregation to impair pinwheel annihilation is preserved as long as Λ_{OD} and Λ_{OP} are roughly similar. In simulations without ocular dominance segregation, $s_o = 0$. Equation (3) defines a dynamics of $z(\mathbf{x})$, $r(\mathbf{x})$ and $o(\mathbf{x})$ in the sense of equation (2). An analysis of its mathematical structure and the algorithm used for its solution are presented in Supplementary Information. In the simulations presented in Figs 3 and 4, $\sigma = 0.89$ and $\eta = 0.005$. Time is measured in units, τ , that correspond to 5×10^4 stimulus applications with $\epsilon = 8.7 \times 10^{-5}$ (see Supplementary Information).

Received 15 May 1997; accepted 19 June 1998.

- Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *J. Physiol.* **160**, 215–243 (1962).
- von der Malsburg, C. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14**, 85–100 (1973).
- Miller, K. D. A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between on- and off-center inputs. *J. Neurosci.* **14**, 409–441 (1994).
- Swindale, N. V. The development of topography in the visual cortex: a review of models. *Network* **7**, 161–247 (1996).
- Movshon, J. A. & Van Sluyters, R. C. Visual neural development. *Annu. Rev. Psychol.* **32**, 447–522 (1981).
- Gödecke, I. & Bonhoeffer, T. Development of identical orientation maps for two eyes without common visual experience. *Nature* **379**, 251–254 (1996).
- Crair, M. C., Gillespie, D. C. & Stryker, M. P. The role of visual experience in the development of columns in cat visual cortex. *Science* **279**, 566–570 (1998).
- Swindale, N. V., Matsubara, J. A. & Cynader, M. S. Surface organization of orientation and direction selectivity in cat area 18. *J. Neurosci.* **7**, 1414–1427 (1987).
- Bonhoeffer, T. & Grünvald, A. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* **353**, 429–431 (1991).
- Maldonado, P. E., Gödecke, I., Gray, C. M. & Bonhoeffer, T. Orientation selectivity in pinwheel centers in cat striate cortex. *Science* **276**, 1551–1555 (1997).
- Albus, K. & Wolf, W. Early post-natal development of neuronal function in the kitten's visual cortex: A laminar analysis. *J. Physiol.* **348**, 153–185 (1984).
- Chapman, B. & Stryker, M. P. Development of orientation selectivity in ferret visual cortex and effects of deprivation. *J. Neurosci.* **13**, 5251–5262 (1993).
- Chapman, B., Stryker, M. P. & Bonhoeffer, T. Development of orientation preference maps in ferret primary visual cortex. *J. Neurosci.* **16**, 6443–6453 (1996).
- Frisch, U. *Turbulence* (Cambridge University Press, Cambridge, 1995).
- Sompolinsky, H. & Shapley, R. New perspectives on the mechanisms for orientation selectivity. *Curr. Opin. Neurobiol.* **7**, 514–522 (1997).
- Bock, G. & Cardew, G. (eds) *Development of the Cerebral Cortex* (Wiley, Chichester, 1995).
- Miller, K. D. Synaptic economics: Competition and cooperation in synaptic plasticity. *Neuron* **17**, 367–370 (1996).
- LeVay, S. & Nelson, S. B. in *Vision and Visual Dysfunction* 266–315 (Macmillan, Houndsmill, 1991).
- Halperin, B. T. in *Physics of Defects, Les Houches, Session XXXV, 1980* (eds Balian, R., Kléman, M. & Poirir, J.-P.) 813–857 (North-Holland, Amsterdam, 1981).

20. Durbin, R. & Mitchinson, G. A dimension reduction framework for understanding cortical maps. *Nature* **343**, 644–647 (1990).
21. Erwin, E., Obermayer, K. & Schulten, K. Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Comp.* **7**, 425–468 (1995).
22. Obermayer, K. & Blasdel, G. G. Singularities in primate orientation maps. *Neural Comp.* **9**, 555–575 (1997).
23. Rao, S. C., Toth, L. J. & Sur, M. Optically imaged maps of orientation preference in primary visual cortex of cats and ferrets. *J. Comp. Neurol.* **387**, 358–370 (1997).
24. Lowel, S., Freeman, B. & Singer, W. Topographic organization of the orientation column system in large flat-mounts of the cat visual cortex: A 2-deoxyglucose study. *J. Comp. Neurol.* **255**, 401–415 (1987).
25. Bonhoeffer, T., Kim, D.-S., Maloney, D., Shoham, D. & Grinvald, A. Optical imaging of the layout of functional domains in area 17 and across the area 17/18 border in cat visual cortex. *Eur. J. Neurosci.* **7**, 1973–1988 (1995).
26. Bosking, W. H., Zhang, Y., Schofield, B. R. & Fitzpatrick, D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* **17**, 2112–2127 (1997).
27. Horton, J. C. & Hocking, D. R. Anatomical demonstration of ocular dominance columns in striate cortex of the squirrel monkey. *J. Neurosci.* **16**, 5510–5522 (1996).
28. Rényi, A. *Probability Theory* (North-Holland, Amsterdam, 1970).
29. German, S. Some averaging and stability results for random differential equations. *SIAM J. Appl. Math.* **36**, 86–105 (1997).
30. Niebur, E. & Wörgötter, F. Design principles of columnar organization in visual cortex. *Neural Comp.* **6**, 602–614 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank S. Löwel, H.-U. Bauer, K. Pawelzik, K. Miller, A. V. M. Herz, H. Dinse, P. Heil and G. Langner for discussion, and S. Löwel, F. Hoffsummer, J. Sandhaas and T. Dresbach for helpful comments on the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft and the Max-Planck-Gesellschaft.

Correspondence and requests for materials should be addressed to F.W. (e-mail: fred@chaos.gwdg.de).

lin-14* regulates the timing of synaptic remodelling in *Caenorhabditis elegans

Steven J. Hallam & Yishi Jin*

Department of Biology, Sinsheimer Laboratories, University of California, Santa Cruz, Santa Cruz, California 95064, USA

In the nematode *Caenorhabditis elegans* six GABAergic motor neurons, known as DDs^{1,2}, remodel their patterns of synaptic connectivity during larval development³. DD remodelling involves a complete reversal of the direction of information flow within nerve processes without marked changes in process morphology. We used a marker localized *in vivo* to DD presynaptic zones to analyse how the timing of DD remodelling is controlled. In wild-type animals, DDs remodel their synaptic outputs within a 3–5-hour period at the end of the first larval stage. We show that the heterochronic gene *lin-14*, which controls the timing of stage-specific cell lineages^{4,5}, regulates the timing of DD synaptic output remodelling. In *lin-14* loss-of-function mutants, DDs remodel precociously. The degree of precocious remodelling is correlated with the level of *lin-14* activity. Expression of *lin-14(+)* in the DDs of *lin-14*-null mutants rescues the precocious remodelling, indicating that *lin-14* can act cell-autonomously. Consistent with this hypothesis, LIN-14 protein levels decrease in the DDs before remodelling. Our observations reveal a role of heterochronic genes in non-dividing cells, and provide an example of cell-autonomous respecification of neuronal connectivity.

Neuronal cells differentiate distinct dendritic branches and axonal processes. This asymmetry, once established, is normally stabilized to maintain the direction of information flow. However, polarity changes occur during insect and amphibian metamorphosis^{6–9} and can be induced in culture¹⁰. The genetic mechanisms controlling when and how neuronal polarity is respecified remain obscure. The phenomenon of DD remodelling in *C. elegans* provides a model in which to explore such mechanisms with single-cell resolution. DD remodelling was first observed using electron-micrographic reconstruction of the *C. elegans* ventral nerve cord³. In first-stage (L1) larvae, DDs innervate ventral body wall muscles,

receiving synaptic inputs on the dorsal side. In later larval and adult (L2–A) stages, DDs innervate dorsal body wall muscles, receiving synaptic inputs on the ventral side³.

To observe DD remodelling *in vivo*, we used a synapse-specific marker, *Punc-25-SNB-GFP*. The promoter of the *unc-25* gene, which encodes glutamic acid decarboxylase (Y.J., E. M. Jorgensen, E. Hartweig, H. R. Horvitz, submitted), is used to drive the expression of a chimaeric protein, SNB-GFP, in which Green Fluorescent Protein (GFP)¹¹ is fused to the C terminus of the *C. elegans* SNB-1 protein, a homologue of synaptobrevin/VAMP¹² (M. Nonet, personal communication). This *unc-25* promoter is active in 23 GABAergic motorneurons, including 4 RMEs, 6 DDs and 13 VDJs (Y.J., *et al.*, submitted). SNB-1 is an integral membrane protein of synaptic vesicles and is localized specifically to presynaptic zones¹². Fluorescence from this GFP marker was seen as punctate clusters along the length of both dorsal and ventral nerve cords (Fig. 1b, e), in regions corresponding to DD and VD presynaptic zones¹³. Moreover, the expression of this marker accurately reflected the patterns of DD synaptic outputs at all developmental stages (Fig. 1a, b, d, e), and in mutants that lack post-embryonic ventral cord neurons (Fig. 1c, f), confirming previous electron micrographic observations³.

We first determined the time period in which DD remodelling occurs by observing the extent of dorsal SNB-GFP localization with respect to developmental time, as indicated by the postembryonic P-cell lineage¹⁴. We found that at 12–13 hours after hatching, coinciding with the third round of division in the P9 lineage, the most anterior DD neuron (DD1) begins to remodel its synaptic outputs, seen as a gradual decrease in ventrally localized SNB-GFP and an appearance of dorsally localized SNB-GFP. The remodelling process continued from DD1 to DD6, in anterior to posterior order along the ventral nerve cord, reaching completion at the end of the L1 moult (Fig. 2). A caveat to these observations on GFP fluorescence is that SNB-GFP might accumulate before visible fluorescence¹⁵. We therefore scored SNB-GFP localization by staining L1 animals with anti-GFP antibodies and visualizing with secondary antibodies. We observed an approximately 1-hour delay between the immunocytochemical detection of dorsal SNB-GFP and GFP fluorescence (not shown). This delay might reflect both the higher sensitivity of immunocytochemical detection and delays in GFP fluorescence after protein synthesis. Despite this delay, the rate and order of DD remodelling were indistinguishable from that observed with GFP fluorescence. We conclude that the expression of the *Punc-25-SNB-GFP* marker accurately reflects DD synaptic remodelling, and that DD remodelling occurs 13–18 hours after hatching, between the third round of P-cell division and the end of the L1 moult.

The *C. elegans* heterochronic gene *lin-14* regulates the timing of stage-specific cell division patterns^{4,5,16}. *lin-14* loss-of-function (*lf*) mutations cause hypodermal, muscle and vulval precursor cells to bypass early-stage patterns and to precociously execute later-stage (L2/L3) patterns of cell division. LIN-14 proteins are expressed in many somatic cell nuclei, including ventral cord neurons during embryogenesis and early larval development¹⁷. However, it is unknown whether *lin-14* functions in neurons. To test whether *lin-14* might regulate the timing of DD remodelling, we examined the expression pattern of *Punc-25-SNB-GFP* in *lin-14(lf)* mutants. Because existing *lin-14(lf)* alleles do not affect the timing of the P-cell nuclear migrations⁵, and DDs do not remodel before the P-cell nuclear migrations in wild-type animals (Fig. 2), we first compared the extent of dorsal SNB-GFP localization before the P-cell nuclear migrations in *lin-14(lf)* and wild-type animals. We found that *lin-14(lf)* mutations cause DDs to remodel precociously (Table 1 and Fig. 3a–e). To define the precise period during which precocious DD remodelling occurs in *lin-14(lf)* animals, we followed the dorsal localization of SNB-GFP with respect to developmental time (Fig. 2). In *lin-14(ma135)* animals, which completely lack

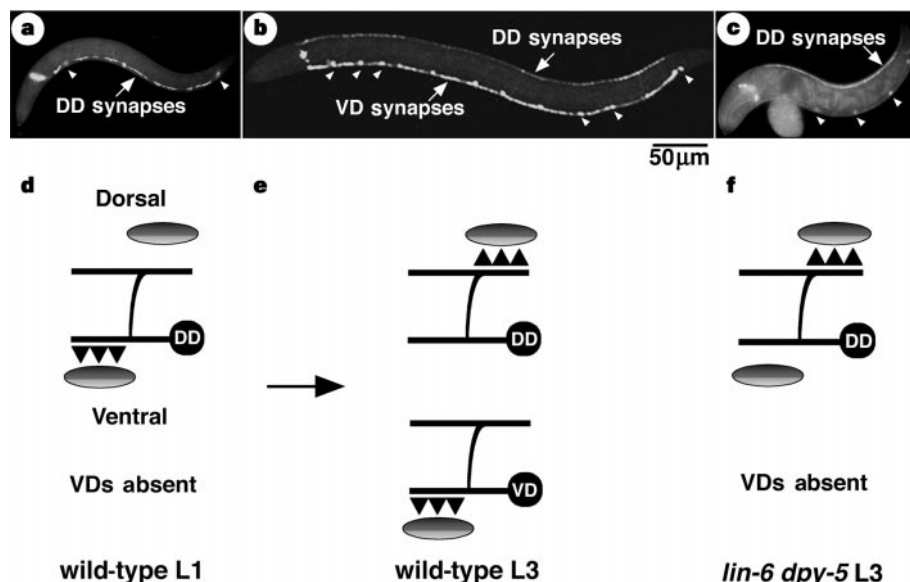


Figure 1 *Punc-25*-SNB-GFP expression reflects the pattern of DD synaptic connectivity. All animals contain *juls1*, the integrated *Punc-25*-SNB-GFP marker. **a**, L1 larva before P-cell nuclear migrations. SNB-GFP is localized on the ventral side, indicating DD innervation of ventral body wall muscles. VDs arise at the end of P-cell divisions in late L1 and are not present at this stage. **b**, L3 larva. The dorsally localized SNB-GFP indicates DD innervation of dorsal body wall muscles, and the ventrally localized SNB-GFP indicates VD innervation of ventral

body wall muscles. **c**, *lin-6* (*e1466*) *dpy-5* (*e61*) L3 larva. In the absence of VDs, DDs remodel normally. In **a–c**, arrows point to SNB-GFP localization; small arrowheads point to some of the DD and VD cell bodies. **d–f**, Schematic representation of DD and VD synaptic outputs in wild-type L1, L3 and *lin-6* (*e1466*) *dpy-5* (*e61*) L3 animals. Arrowheads depict neuromuscular junctions, shaded ovals are body wall muscles, closed circles are neuronal cell bodies, and solid lines represent nerve processes.

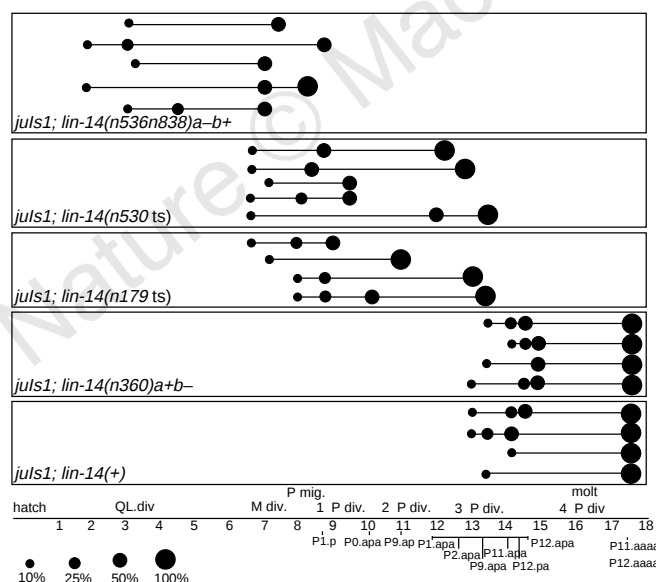


Figure 2 Timing of DD synaptic remodelling. The horizontal axis represents developmental time determined by L1-stage-specific events including QL division (QL div.), M cell division (M div.), P-cell migrations (P mig.) and rounds of division (1°, 2°, 3° and 4° div.). Each horizontal line represents one animal followed. Circles represent individual time points observed for each animal. Circle size corresponds to the extent of SNB-GFP localization along the dorsal cord, ranging from ≤10% (initial detection) to 100% (complete). Owing to the presence of newly formed synapses between VDs and ventral body wall muscle in late L1, the completeness of DD synaptic remodelling could not be determined for *juls1*; *lin-14* (+) and *juls1*; *lin-14* (*n360*).

lin-14 activity¹⁸, DDs exhibited their adult patterns of synaptic connectivity upon hatching (Fig. 3e); we have not determined whether DDs ever exhibit the L1 larval patterns of connectivity in *lin-14* (*ma135*) embryos. In animals bearing the temperature-sensitive *lin-14* (*n179ts*) and *lin-14* (*n530ts*) mutations, both of which partly decrease *lin-14* activity, DD remodelling began 5–7 hours after hatching, reaching completion 10–12 hours after hatching (Figs 2, 3b). Although the onset of DD remodelling was precocious in these *lin-14* (*lf*) mutants, the rate of remodelling was similar to that in wild-type animals. From these observations we conclude that *lin-14* regulates the onset of DD remodelling and that the degree of precocious remodelling is correlated with the level of *lin-14* activity.

Previous genetic analysis identified two *lin-14* activities: *lin-14a*, required for L1–stage-specific cell lineages, and *lin-14b*, required for L2–stage-specific cell lineages^{5,16}. To determine which of these two *lin-14* activities regulates DD remodelling, we followed the dorsal localization of SNB-GFP with respect to developmental time in animals bearing either *lin-14* (*n536n838*), a mutation that disrupts *lin-14a* activity alone⁵, or *lin-14* (*n360*), a mutation that disrupts *lin-14b* activity alone⁵ (Table 1 and Fig. 2). In *lin-14* (*n536n838*) mutants, DDs began to remodel 2–4 hours after hatching, reaching completion 7–9 hours after hatching, coinciding with the second round of P-cell divisions (Figs 2, 3c). The effect of *lin-14* (*n536n838*) on DD remodelling is less severe than *lin-14* (*ma135*), perhaps because *lin-14* (*n536n838*) does not completely eliminate *lin-14a* activity⁵. In *lin-14* (*n360*) animals the timing of DD remodelling was indistinguishable from that of wild-type animals (Figs. 2, 3d). Thus the timing of DD synaptic remodelling is specified by *lin-14a* activity.

We next examined how the level of LIN-14 protein changes in DDs before and during the remodelling process. We found that in wild-type animals, coinciding with the onset of P-cell nuclear migrations, LIN-14 expression was downregulated in the DDs

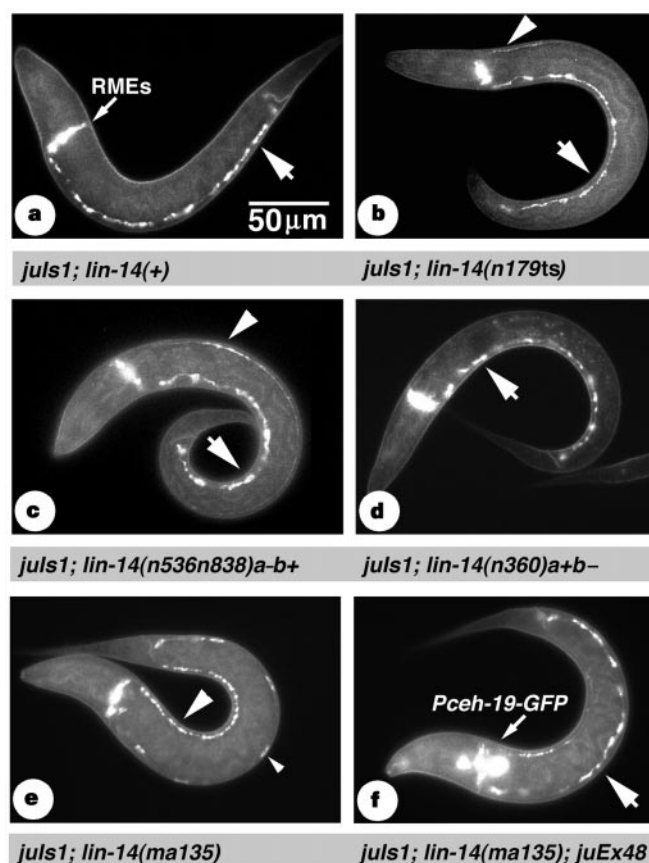


Figure 3 Effects of *lin-14* on DD synaptic remodelling. Animals of all genotypes are L1 larvae 2–7 h after hatching, before P-cell nuclear migrations. **a**, *juls1; lin-14(+)*: SNB-GFP is localized exclusively on the ventral side. **b**, *juls1; lin-14(n179ts)*: the dorsally localized SNB-GFP indicates precocious DD remodelling. The range is estimated as 20%. **c**, *juls1; lin-14(n536n838)a-b+*: the range of precocious DD remodelling is ~60%. **d**, *juls1; lin-14(n360)a-b-*: no precocious DD remodelling. **e**, *juls1; lin-14(ma135)*: DDs have remodelled completely. **f**, *juls1; lin-14(ma135); juEx48*: expression of *lin-14(+)* in DDs rescues the precocious remodelling. Owing to mosaicism of the extrachromosomal array, not all DDs show rescue. Large arrows mark ventral SNB-GFP localization; large arrowheads depict precocious DD remodelling; small arrowhead in **e** marks a DD neuron cell body; thin arrows in **a** and **f** represent non-D GABAergic neurons and the *Pceh-19* GFP marker. In all photographs anterior is to the left, and dorsal is up.

(Fig. 4a–c, d–f). Interestingly, LIN-14 levels in the DDs increased transiently during the first round of P-cell division before decreasing in later larval and adult stages (Fig. 4h–i). The *lin-14* locus encodes several isoforms that share a large common C-terminal region¹⁹; the LIN-14 antibody we used was raised against the common region¹⁷. The transient decrease of LIN-14 in mid-L1 followed by a transient increase in late L1 might reflect either the regulation of all LIN-14 isoforms or changes in the level of specific LIN-14 isoforms. To further demonstrate how *lin-14* activity is required for DD remodelling, we expressed wild-type *lin-14* specifically in the DDs of *lin-14(ma135)* animals (see Methods). We observed complete rescue of precocious remodelling in *lin-14(ma135)* L1 larvae upon DD-specific expression of *lin-14(+)* (Fig. 3f). Moreover, the onset of DD remodelling in these transgenic animals was significantly delayed, with DDs in L2 animals retaining synaptic outputs to the ventral body wall muscles (not shown). We have not determined precisely the onset time of DD remodelling in these rescued animals. None the less, our results suggest that LIN-14 can act cell-autonomously to regulate the timing of DD remodelling.

Although the mode of LIN-14 action at the molecular level is unknown, previous genetic and biochemical analyses have shown that high levels of LIN-14 protein promote L1-stage-specific patterns of cell division, whereas reduced levels of LIN-14 protein promote L2- and L3-stage-specific patterns of cell divisions^{4,5,16,18,20}. In all of these cases, *lin-14* functions in mitotically competent cells, either by regulating the particular pattern of cell division that they generate or by controlling when they divide. In contrast, we observe that reductions in *lin-14* activity within the DDs cause precocious synaptic remodelling in the absence of cell division, providing the first evidence that *lin-14* can regulate the timing of events in post-mitotic cells.

Examples of neuronal remodelling are widespread, from simple addition and elimination of synapses according to usage and age^{21–26}, to major nervous system reorganization during metamorphosis^{6–9}. The phenomenon of DD remodelling is, to our knowledge, unique in that a complete reversal of the direction of information flow within nerve processes occurs in the absence of marked changes in process morphology. To remodel their synapses, DD neurons must either reroute or eliminate existing molecules, and might initiate *de novo* synthesis and transport. We have shown that the heterochronic gene *lin-14* can control this complex event cell-autonomously. Understanding how *lin-14* controls the reorganization of specialized synaptic components within a neuron might provide insight into how neuronal polarity and connectivity are determined and respecified. □

Table 1 Precocious DD remodelling in *lin-14(lf)* mutants

Genotype	Temperature (°C)	Number of animals with dorsal SNB-GFP localization before P-cell nuclear migrations	Range of dorsal SNB-GFP localization (%)
<i>juls1; lin-14(+)</i>	20	1/57 (2%)	≤10*
	22.5	1/37 (3%)	≤10*
	25	1/36 (3%)	≤10*
<i>juls1; lin-14(n179ts)</i>	22.5	15/29 (52%)	~10
	25	65/83 (65%)	~10
<i>juls1; lin-14(n530ts)</i>	22.5	30/43 (70%)	10–25
	25	30/32 (94%)	10–25
<i>juls1; lin-14(n360)a-b-</i>	22.5	1/43 (2%)	≤10*
	25	1/25 (4%)	≤10*
<i>juls1; lin-14(n536n838)a-b+</i>	22.5	42/50 (84%)	10–50
	25	44/45 (98%)	10–50
<i>juls1; lin-14(ma135)</i>	20	37/37 (100%)	100
	22.5	21/21 (100%)	100

juls1 animals were staged at 20, 22.5 and 25 °C to assess the effect of temperature on DD remodelling. No temperature dependence was observed in this background. In contrast, both *lin-14(n179ts)* and *lin-14(n530ts)* alleles showed an increase in the penetrance of precocious DD remodelling at the restrictive temperature (25 °C). Early L1 larvae were identified on the basis of undescended P-cell nuclei¹⁴. P-cells are precursors for postembryonic ventral cord neurons. On hatching, they are located in a ventro-lateral position; 7–9 h after hatching, P-cell nuclei descend into the ventral nerve cord, and subsequently P-cells undergo four rounds of cell division. The timing of P-cell nuclear migration and division is invariant. Data represent the sum of at least three separate staging events for each allele on separate days. Range was determined as the estimated percentage extent of dorsal SNB-GFP localization along the length of the dorsal cord, ≤10% representing initial detection and 100% representing completion. An asterisk indicates that only one animal was observed with extremely faint anterior dorsal SNB-GFP localization. When possible, animals showing this pattern were recovered and observed at later times to see whether remodelling continued. In most cases the presence of this dorsal SNB-GFP could not be detected until the onset of the third round of P-cell division and therefore those cases were not scored positive.

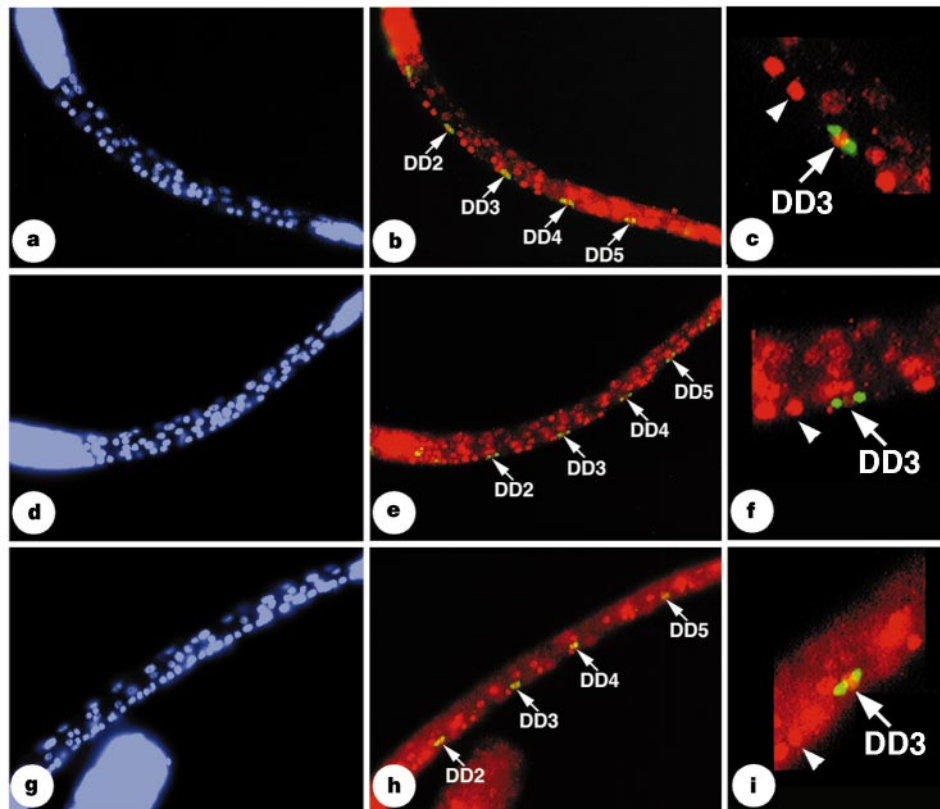


Figure 4 Changes in LIN-14 protein levels within the DDs precede the onset of synaptic remodelling. All animals are L1 larvae bearing *juIs7*, the *Punc-25-GFP* marker. **a, d, g**, DAPI (4',6-diamidino-2-phenylindole) staining to show the age of the animal, as determined by the position of P-cell nuclei and the number of nuclei in the ventral nerve cord. **b, c, e, f, h, i**, LIN-14 protein levels (red) in DDs (green). DDs (arrows) are identified by GFP fluorescence from *juIs7*. **b, c**, On hatching,

LIN-14 in DDs is high and equal in intensity to other ventral cord neurons (arrowhead). **e, f**, 7–9 h after hatching, coinciding with the onset of P-cell nuclear migrations, LIN-14 in DDs decreases compared to other ventral cord neurons (arrowhead). **h, i**, 9–11 h after hatching, coinciding with the first round of P-cell division, LIN-14 in DDs increases transiently before decreasing in later larval and adult stages. In all photographs anterior is to the left.

Methods

Genetics. *C. elegans* strains were maintained as described²⁷. Our reference strain was CZ333, containing *juIs1* IV, an insertion of the [*Punc-25-SNB-GFP*; *lin-15(+)*] extrachromosomal array in *C. elegans* var. Bristol strain N2. CZ333 was generated as follows. First, the *Punc-25-SNB-GFP* DNA (pSC314) was injected into the germline of *lin-15(n765ts)* mutants along with *pln-15EK*, a plasmid containing the *lin-15* gene²⁸. Non-Muv (multivulva) animals were selected at 22.5°C and screened for expression of the *Punc-25-SNB-GFP* marker. Second, the stably transmitted extrachromosomal array was integrated into the genome by X-ray mutagenesis. Four integrants were isolated after screening 1200 X-ray-mutagenized F₁s. Integrants were backcrossed against N2 at least five times to remove *lin-15(n765ts)*. All integrants gave similar GFP expression patterns with varying degrees of fluorescence intensity; here we present data collected with *juIs1*. Mutant *lin-14* alleles are as described^{4,5,18}. CZ520 *juIs7* is an insertion of the [*Punc-25-GFP*; *lin-15(+)*] array. This *Punc-25-GFP* marker contains GFP fused in-frame to the amino-terminus of the UNC-25 protein, and expresses GFP in the perinuclear cytoplasm of all GABAergic neurons (Y.J., *et al.*, submitted). All strains containing *juIs1* and *lin-14(lf)* mutations, with the exception of *lin-14(ma135)*, were created by crossing *juIs1* males with *lin-14(lf)* mutant hermaphrodites and establishing strains homozygous for both *juIs1* and *lin-14(lf)*. *lin-14(ma135)* was maintained in a balanced strain using the balancer chromosome *szT1 I:X* after mating with *juIs1* males.

Staging and lineage analysis. Synchronous populations of early L1 animals were collected by lysing gravid adults in hypochlorite. Embryos were allowed to develop at 20, 22.5 or 25°C for an average of 16 h before observation under Nomarski/fluorescence optics on a Zeiss Axioskop microscope. All lineage analysis was conducted at room temperature (21–23°C) as described¹⁴.

For photography, mixed-stage worm populations were collected, fixed in 2% paraformaldehyde for 15 min at room temperature, and washed three times in M9 for 15 min before mounting (DABCO in 50% glycerol; Aldrich, Wisconsin). GFP fluorescence was observed using an HQ FITC-LP filter, with excitation at 455–500 nm and emission at 510+ nm (Chroma, Vermont) on a Zeiss Axioskop microscope.

Molecular biology and germline transformation. The *Punc-25-lin-14(+)* clone (pCZ124) was constructed by inserting a 1.5-kb *SphI*–*BglII* *unc-25* promoter at the 5' end of a 5.8 kb *lin-14* genomic DNA fragment. This *lin-14* DNA contains sequences from exon 4 to exon 13, which includes the common region for all LIN-14 isoforms, and the 3' untranslated regulatory region. It is sufficient to rescue all *lin-14* mutant phenotypes (Y. Hong and V. Ambros, personal communication). pCZ124 DNA (50 ng μ l⁻¹) was injected into CZ355 *+szT1 I:X*; *lin-14(ma135)/szT1 I:X*, along with pCZ125 (50 ng μ l⁻¹), a reporter construct in which GFP is driven by the promoter of the *ceh-19* gene²⁹ and is expressed in several amphid neurons (Y.J., unpublished results). Transformants were identified on the basis of pCZ125 expression. Several stably transmitting lines, including CZ747 *juIs1* IV; *+szT1 I:X*; *lin-14(ma135)/szT1 I:X*; *juEx48*, were established. DD-specific expression of LIN-14 was confirmed by staining animals with anti-LIN-14 antibodies (not shown).

Immunocytochemistry. Mixed-stage populations of CZ333 *juIs1* and CZ520 *juIs7* were collected, fixed and permeabilized, then incubated with affinity-purified anti-LIN-14 (ref. 17) or anti-GFP sera (Clontech), and visualized by fluorescein isothiocyanate or Texas-red-conjugated secondary antibodies as described¹⁷.

Received 7 May; accepted 26 June 1998.

1. McIntire, S. L., Jorgensen, E., Kaplan, J. & Horvitz, H. R. The GABAergic nervous system of *Caenorhabditis elegans*. *Nature* **364**, 337–341 (1993).

- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
- White, J. G., Albertson, D. G. & Anness, M. A. R. Connectivity changes in a class of motoneurons during the development of a nematode. *Nature* **271**, 764–766 (1978).
- Ambros, V. & Horvitz, H. R. Heterochronic mutants of the nematode *Caenorhabditis elegans*. *Science* **226**, 409–416 (1984).
- Ambros, V. & Horvitz, H. R. The *lin-14* locus of *Caenorhabditis elegans* controls the time of expression of specific postembryonic developmental events. *Genes Dev.* **1**, 398–414 (1987).
- Harris, W. A. Neurometamorphosis. *J. Neurobiol.* **21**, 953–957 (1990).
- Truman, J. Metamorphosis of the central nervous system of *Drosophila*. *J. Neurobiol.* **21**, 1072–1084 (1990).
- Levine, R. B., Morton, D. B. & Restifo, L. L. Remodeling of the insect nervous system. *Curr. Opin. Neurobiol.* **5**, 28–35 (1995).
- Alley, K. E. Retrofitting larval neuromuscular circuits in the metamorphosing frog. *J. Neurobiol.* **21**, 1092–1107 (1990).
- Dotti, C. G. & Banker, G. A. Experimentally induced alterations in the polarity of developing neurons. *Nature* **330**, 254–256 (1987).
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
- Nonet, M. L., Saifee, O., Zhao, H., Rand, J. B. & Wei, L. Synaptic transmission deficits in *Caenorhabditis elegans* synaptobrevin mutants. *J. Neurosci.* **18**, 70–80 (1998).
- Jorgensen, E. M. et al. Defective recycling of synaptic vesicles in synaptotagmin mutants of *Caenorhabditis elegans*. *Nature* **378**, 196–199 (1995).
- Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
- Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autooxidation of green fluorescent protein. *Proc. Natl Acad. Sci. USA* **91**, 12501–12504 (1994).
- Ambros, V. & Moss, E. G. Heterochronic genes and the temporal control of *C. elegans* development. *Trends Genet.* **10**, 123–127 (1994).
- Ruvkun, G. & Giusto, J. The *Caenorhabditis elegans* heterochronic gene *lin-14* encodes a nuclear protein that forms a temporal developmental switch. *Nature* **338**, 313–319 (1989).
- Euling, S. & Ambros, V. Heterochronic genes control cell cycle progress and developmental competence of *C. elegans* vulval precursor cells. *Cell* **84**, 667–676 (1996).
- Wightman, B., Bürglin, T. R., Gatto, J., Arasu, P. & Ruvkun, G. Negative regulatory sequences in the *lin-14* 3′-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. *Genes Dev.* **5**, 1813–1824 (1991).
- Wightman, B., Ha, I. & Ruvkun, G. Post-transcriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862 (1993).
- Purves, D. & Hadley, R. D. Changes in the dendritic branching of adult mammalian neurones revealed by repeated imaging *in situ*. *Nature* **315**, 404–406 (1985).
- Purves, D. & Lichtman, J. W. Elimination of synapses in the developing nervous system. *Science* **210**, 153–157 (1994).
- Wigston, D. J. Repeated *in vivo* visualization of neuromuscular junctions in adult mouse lateral gastrocnemius. *J. Neurosci.* **10**, 1753–1761 (1990).
- Chen, L., Folsom, D. B. & Ko, C.-P. The remodelling of synaptic extracellular matrix and its dynamic relationship with nerve terminals at living frog neuromuscular junctions. *J. Neurosci.* **11**, 2920–2930 (1991).
- Bailey, C. H. & Kandel, E. R. Structural changes accompanying memory storage. *Annu. Rev. Physiol.* **55**, 397–426 (1993).
- Goodman, C. S. & Shatz, C. J. Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* **72**, 77–98 (1993).
- Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
- Clark, S. G., Lu, X.-W. & Horvitz, H. R. The *Caenorhabditis elegans* locus *lin-15*, a negative regulator of a tyrosine kinase signaling pathway, encodes two different proteins. *Genetics* **137**, 987–997 (1994).
- Naito, M., Kohara, Y. & Kurosawa, Y. Identification of a homeobox-containing gene located between *lin-45* and *unc-24* on chromosome IV in the nematode *Caenorhabditis elegans*. *Nucleic Acids Res.* **20**, 2967–2969 (1992).

Acknowledgements. We thank M. Nonet for the SNB–GFP DNA; V. Ambros and Y. Hong for *lin-14* alleles and *lin-14*(+) genomic DNA; G. Ruvkun, B. Reinhart and F. Slack for anti-LIN-14 antibodies; and A. Chisholm, A. Zahler, J. Sisson, V. Ambros, M. Nonet, M. Zhen, I. Chin-Sang and members of the Jin and Chisholm labs for helpful discussions and comments. Some strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH Center for Research Resources. S.J.H. was supported by a Predoctoral GAANN fellowship. Y.J. is an Alfred P. Sloan Research Fellow.

Correspondence and requests for materials should be addressed to Y.J. (e-mail: jin@biology.ucsc.edu).

Three-dimensional segregation of supramolecular activation clusters in T cells

Colin R. F. Monks*, Benjamin A. Freiberg†‡, Hannah Kupfer*, Noah Siciak* & Abraham Kupfer*†‡

* Division of Basic Science, Department of Pediatrics, National Jewish Medical and Research Center, 1400 Jackson Street, Denver, Colorado 80206, USA
 † Departments of ‡ Immunology and ‡ Cellular and Structural Biology, University of Colorado Health Sciences Center, Denver, Colorado 80262, USA

Activation of T cells by antigen-presenting cells (APCs) depends on the complex integration of signals that are delivered by multiple antigen receptors. Most receptor-proximal activation events in T cells^{1,2} were identified using multivalent anti-receptor antibodies, eliminating the need to use the more complex APCs.

As the physiological membrane-associated ligands on the APC and the activating antibodies probably trigger the same biochemical pathways, it is unknown why the antibodies, even at saturating concentrations, fail to trigger some of the physiological T-cell responses³. Here we study, at the level of the single cell, the responses of T cells to native ligands. We used a digital imaging system and analysed the three-dimensional distribution of receptors and intracellular proteins that cluster at the contacts between T cells and APCs during antigen-specific interactions^{3,4}. Surprisingly, instead of showing uniform oligomerization, these proteins clustered into segregated three-dimensional domains within the cell contacts. The antigen-specific formation of these new, spatially segregated supramolecular activation clusters may generate appropriate physiological responses and may explain the high sensitivity of the T cells to antigen.

Engaged and activated lymphocyte-bound receptors, together with their associated proteins, can be identified and studied, at the single-cell level, by their localized clustering at the T-cell–APC contact^{3–6}. Cloned D10 T cells, specific for the antigens conalbumin and IAK, were mixed with APCs that had been pulsed with conalbumin (500 µg ml^{−1}), and, 30 min later, the cell mixtures were fixed and doubly immunofluorescently labelled with antibodies specific for talin and for protein kinase C (PKC)-θ (Fig. 1). Sets of 46 serial optical sections, 0.2 µm apart along the z axis, were recorded for each of the immunofluorescent labels. The three-dimensional image sets, which span the entire boundaries of the cell conjugates, were computationally processed using deconvolution software to eliminate the out-of-focus haze, thereby improving the spatial resolution and fidelity of the data⁷.

Figure 1a, b shows the labelling for talin and PKC-θ in one of the deconvolved optical sections, through the middle of the cell conjugate. Such single two-dimensional images show that both talin and PKC-θ clustered at the cell contact, as reported previously^{3,8}, but they also indicated some differential localization. Surprisingly, the reconstructed three-dimensional views of the cell contacts showed that these proteins were present in two distinct, non-overlapping domains (Fig. 1d, e). Talin was enriched in most of the contact area, but was absent from a smaller central region (Fig. 1d). In contrast, PKC-θ was enriched in only a small contact area (Fig. 1e) from which talin was largely excluded (Fig. 1f). The same contact domains were also seen when other anti-talin and PKC-θ antibodies

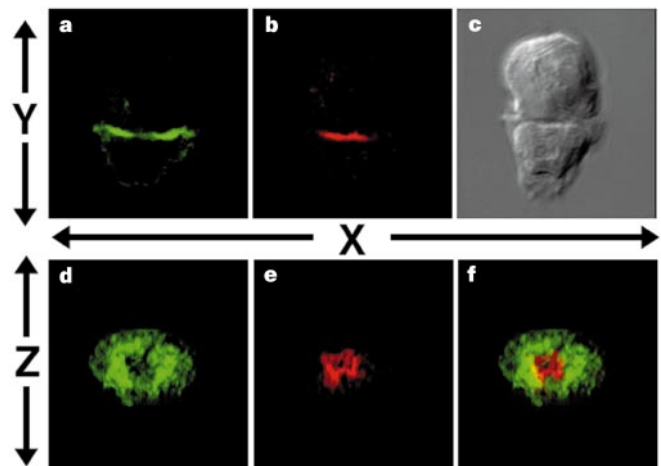


Figure 1 The localization of talin and PKC-θ at the three-dimensionally reconstructed T-cell–APC contact site. The antigen-specific D10–CH12 cell conjugates were labelled with guinea-pig antibodies specific for talin (a, d) and rabbit antibodies specific for PKC-θ (b, e). a, b, A single 2D optical section (along the x–y axis) in the centre of the cell conjugate whose Nomarski image is shown in c. d–f, The 3D view of the entire cell contact (along the x–z axis). f, Both talin (green) and PKC-θ (red) are shown at the cell contact. Note that the two proteins are present in two distinct domains.

were used and when the fluorescent labels for talin and PKC- θ were switched (data not shown). Moreover, similar three-dimensional contact domains were seen in control experiments in which the T-cell-APC conjugates were labelled singly with either anti-talin or anti-PKC- θ antibodies (data not shown). None of these domains were seen in antigen-nonspecific conjugates, for which neither talin nor PKC- θ was enriched in the contacts.

The activation-induced segregation of talin and PKC- θ indicated that the molecular composition of the cytoplasmic side of the contact is not uniform. To determine whether clustered receptors also spatially segregate in the T-cell-APC contact sites, we analysed the three-dimensional distribution of two receptors, LFA-1 and the T-cell antigen receptor (TCR). LFA-1 associates with talin, but not with PKC- θ , in activated T cells^{3,4,9}. Activation of the TCR by the APC is required for the translocation of both PKC- θ and talin,

but antibody co-capping experiments failed to show an association between the TCR and either talin or PKC- θ ^{3,4,9}. We conjugated T cells from AD10-TCR transgenic mice with CH12 cells, which had been pulsed with the antigen pigeon cytochrome C (PCC) (50 $\mu\text{g ml}^{-1}$). The cell mixtures were doubly labelled with anti-LFA1 and either anti-talin or anti-PKC- θ antibodies (Fig. 2). The immunofluorescent images show that LFA-1, talin and PKC- θ cluster at the cell contact in antigen-specific conjugates (Fig. 2b, c, h, i) but not in antigen-nonspecific conjugates (Fig. 2n, o). The three-dimensional reconstructed views of the cell contacts showed that LFA-1 and talin were enriched in most of the contact, and both were excluded from the same central contact area (Fig. 2d-f). In contrast, LFA-1 and PKC- θ were segregated into two different contact domains (Fig. 2j-l).

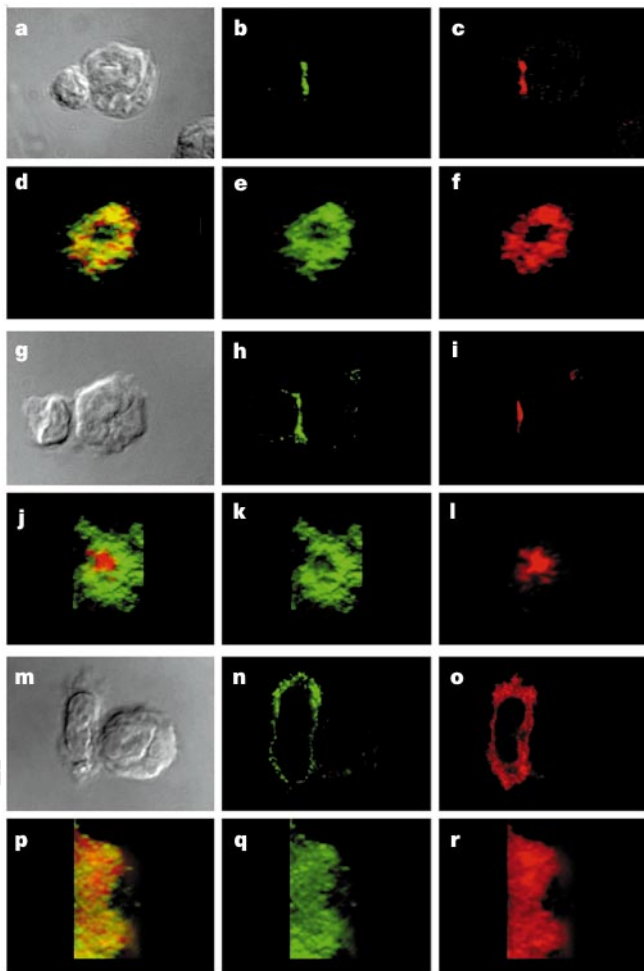


Figure 2 The 3D redistribution of LFA-1 in antigen-specific T-APC conjugates. In Figs 2, 3, 5, 7 and 8, each cell conjugate is shown in two rows of three panels. The top row of each pair of rows starts with a Nomarski image. The next two panels show each of the two labels in a single optical section along the x - y axis. The bottom row of each pair starts with a panel showing both labels in the 3D view of the entire contact area (x - z axis). The next two panels show each of the corresponding labels in the 3D view of the contact. The 3D views were magnified by 50% to show more detail. Antigen-specific (a-l) and antigen-nonspecific (m-r) conjugates of AD10-CH12 cells are shown. The cells were labelled with rat antibodies specific for LFA-1 (in green, b, e, h, k, n, q), guinea-pig antibodies against talin (in red, c, f, i, l, o, r) or rabbit antibodies specific for PKC- θ (in red, j, m). Note that LFA-1 is enriched in most of the contact, like talin, but is excluded from PKC- θ . None of these proteins cluster and organize in the antigen-nonspecific conjugates.

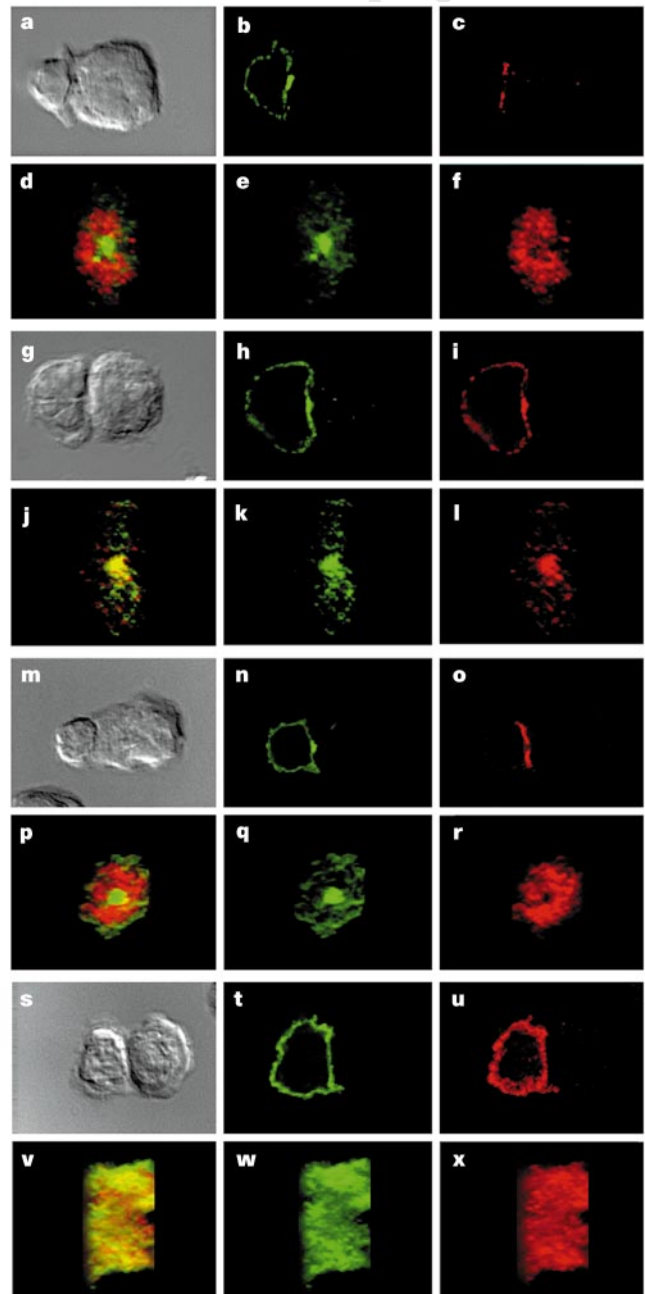


Figure 3 The 3D redistribution of TCR-CD3 in antigen-specific (a-r) but not in antigen-nonspecific (s-x) T-cell-APC conjugates. D10-CH12 (a-l) and AD10-CH12 (m-x) cell conjugates were labelled with anti-CD3 antibodies (in green, b, e, h, k, n, q, t, w) and any of the following antibodies (in red): anti-LFA1 (c, f), anti-V β 8 chain of the TCR (i, l) or anti-talin (o, r, u, x). See Fig. 2 legend for layout of panels.

The formation of these three-dimensional domains depended on recognition of specific antigen. In antigen-nonspecific conjugates, LFA-1 and talin remained randomly distributed (Fig. 2p–r).

We determined the spatial distribution of the TCR–CD3 complex in antigen-specific AD10–CH12 and D10–CH12 cell conjugates (Fig. 3). Clusters of CD3 and LFA-1 were observed in the contact area (Fig. 3a–f). There were, however, significant quantitative and qualitative differences between the clusters of LFA-1 and CD3. Quantitative three-dimensional analysis of the entire cell membranes showed that $78 \pm 9\%$ (mean $\pm$ s.d.) ($n = 20$) of total surface LFA-1 was present at the cell contacts, and only $42 \pm 9\%$ ($n = 20$) of the TCR was in the contact (Fig. 4a). The fraction of the TCR in the contact was only slightly higher than that of the class I major histocompatibility complex (MHC) protein H2k, a membrane protein that did not cluster ($36 \pm 5\%$; $n = 10$) (Figs 4a and 5o, r). The contact may contain both ligand-bound and unbound receptors. Receptor ligation at the cell contact results in a localized enrichment that is proportional to the extent of binding^{4,6,10}. We determined the fraction of clustered (that is, bound) LFA-1 and TCR by quantifying contact voxels whose intensity was higher than the intensity of the voxels in the rest of the T-cell membrane. In this analysis, only a small fraction of total TCR ($6 \pm 3\%$) appeared to be ligand-bound in the contact. In contrast, most LFA-1 ($78 \pm 9\%$) appeared to be ligand-bound (Fig. 4b). This difference concurred with the expected relative concentrations of the cell-adhesion molecule ICAM and specific peptide–IAk (IAk is a class II MHC protein) complexes¹¹, the ligands for the LFA-1 and TCR, respectively. In addition, although clustered LFA-1 was present along most of the contact area, the clustered TCR was confined to only $6.2 \pm 2.4\%$ of the contact area. To estimate the frequency of TCR clustering among all T-cell–APC couples, we used Nomarski optics to identify cell conjugates and blindly imaged their TCR labelling. Clustered TCR was detected in 28 of such 29 randomly selected cell conjugates.

Three-dimensional analysis of the cell contacts showed that clustered LFA-1 and TCR were spatially segregated from each other (Fig. 3d–f). Similar three-dimensional domains were seen in control experiments in which the cells were either singly labelled for only one of the receptors or doubly labelled with other antibody combinations. These results indicate that the mutual exclusion of the two receptors reflected their distinct spatial distribution and was

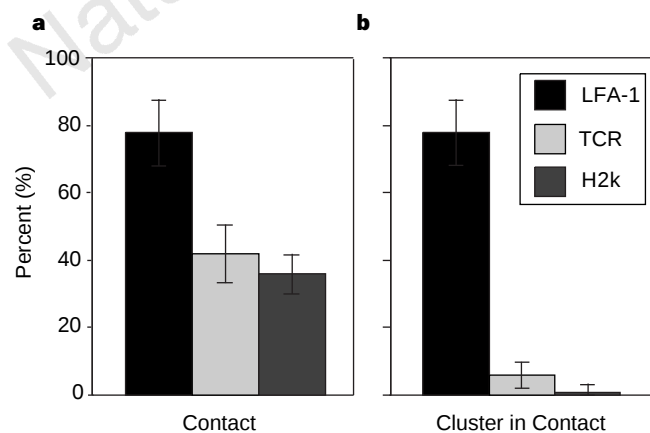


Figure 4 Quantitative analysis of the fractional distribution of receptors for the T-cell–APC contact site. Conjugates of D10 and CH12 cells, which were pulsed with $500 \mu\text{g ml}^{-1}$ conalbumin, were singly labelled with anti-LFA1, anti-V β 8 chain of the TCR or anti-H2k antibodies. 3D images of cell couples were recorded and the localized surface expression of the different receptors was quantified. **a**, The labelling in the entire cell contact expressed as a percentage of total T-cell labelling. **b**, The clustered (that is, engaged) receptors at the contact as a percentage of total T-cell labelling.

not due to an inability to label two antibodies in the same domain. The clustering of the TCR was further confirmed by doubly labelling D10–CH12 cells with antibodies directed against CD3– ϵ and V β 8–TCR (Fig. 3g–l), as well as by labelling the cytoplasmic tail of the ζ -chain of CD3 (not shown). The spatial segregation of the clustered TCR was detected in several different antigen-specific T-cell clones (as in Fig. 3) and, significantly, in *ex vivo* T cells from TCR-transgenic mice (Fig. 3a–f). The TCR did not cluster and did not form a centrally focused domain in antigen-nonspecific conjugates of T-helper (T_H) cells and APCs (Fig. 3t, v, w).

In extra control experiments, AD10–CH12 conjugates were doubly labelled with anti-CD3 antibodies and with a murine antibody specific for H2k (Fig. 5m–r). H2k is a membrane glycoprotein that is not expected to be involved in this interaction, as the TCRs of the AD10 T cells recognize peptide that is bound to class II, not class I, MHC. Although the TCR–CD3 was focused in the contact (Fig. 5n, p, q) the labelling for H2k remained uniform and was neither enriched nor depleted from any domain (Fig. 5o, p, r). These results indicate that the segregation of membrane proteins into contact domains is a specific and regulated process, and is not caused by the passive sorting of proteins. Thus, the three-dimensional analysis of the entire site of interaction between T_H cells and APCs shows, surprisingly, that receptors and intracellular proteins that are

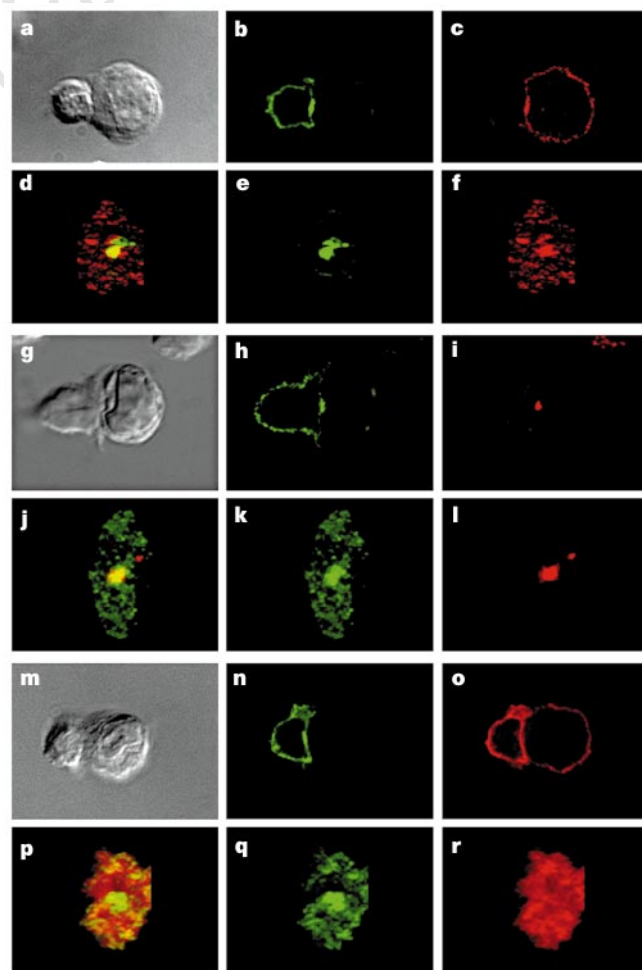


Figure 5 The 3D redistribution of TCR–CD3 and class II MHC in antigen-specific T-cell–APC conjugates. AD10–CH12 (**a–f** and **m–r**) and D10–CH12 (**g–l**) conjugates were labelled with anti-CD3 antibody (in green, **b, e, h, k, n, q**) and with one of the following antibodies (in red): anti-IE α (**c, f**), anti-IA α (**i, l**) and anti-H2k (**o, r**). Note that class II antigens co-cluster with the TCR but that class I antigens do not co-cluster. See Fig. 2 legend for panel layout.

involved in physiological T-cell activation are organized into distinct spatial domains. We name these three-dimensional contact domains 'supramolecular activation clusters' (SMACs), which include the peripheral SMAC (p-SMAC) and the central TCR-SMAC (c-SMAC) (Fig. 6).

Because the SMACs are seen only in antigen-specific conjugates, it is likely that the c-SMAC is the site of TCR ligation. To further test this idea, we determined whether class II MHC, TCR ligand, clusters in the APC, mirroring the c-SMAC in the T cell. As the CH12 cells express both IAK and IEK class II MHC, we used AD10 T cells that recognize PCC peptide bound to IEK and D10 T cells that recognize conalbumin peptide bound to IAK. 14-4-4, a murine monoclonal antibody specific for IEK, and a rabbit antibody directed against the cytoplasmic tail of IA- α (ref. 12) were used to selectively localize IEK and IAK, respectively. Clustering of surface IEK (Fig. 5c) was detected at the contact of AD10-CH12 conjugates that were doubly labelled with anti-CD3 and anti-IEK antibodies (Fig. 5a-f). Remarkably, the clustered IEK on the APC mirrored the c-SMACs on the T cell (Fig. 5d-f). Unlike IEK, labelled IAK was not clustered at the AD10-CH12 contact area (results not shown). The reverse results were seen in D10-CH12 cells. In these conjugates, IAK, but not IEK, was clustered on the CH12 cells (Fig. 5g-l). Again, the IAK cluster mirrored the c-SMAC, indicating that the clustered TCR-CD3 complexes at the c-SMAC are probably bound to their ligands.

To determine whether the SMACs are important for T-cell activation, CH12 cells were treated with CA37 ($5 \mu\text{g ml}^{-1}$), an agonist peptide, or with either E8A or E8T ($5 \mu\text{g ml}^{-1}$), which are altered peptides containing a single amino-acid substitution at position 8 (ref. 13). Although the agonist peptide causes extensive T-cell proliferation and cytokine production, the altered peptides function as antagonist peptides for T cells derived from D10-TCR transgenic mice and do not trigger cytokine production or T-cell proliferation (for a complete description of these peptides, see ref. 13). The T-cell-APC conjugates were doubly labelled with anti-

CD3 antibody, to detect the c-SMAC, and anti-talin antibody, to visualize p-SMAC (Fig. 7). In the agonist-induced cell conjugates, both talin and CD3 clustered at the cell contact and both p- and c-SMACs were formed (Fig. 7a-f). Talin and CD3 also clustered at the cell contacts in the altered-peptide-induced conjugates (Fig. 7h, i), indicating that the TCR was engaged and that activation signals for talin translocation were delivered. This is in agreement with previous studies showing that activation of talin clustering can be

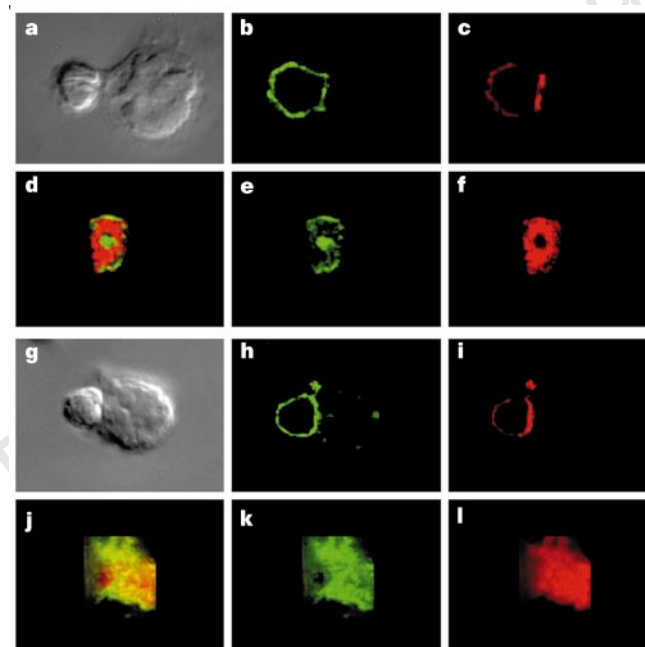


Figure 7 Agonist peptides and not altered peptides cause the formation of SMACs. Cell conjugates containing T cells from D10-TCR transgenic mice and CH12 cells that were pulsed with either CA37, an agonist peptide (a-f), or E8A, an antagonistic peptide (g-l), were labelled with anti-CD3 antibody (in green, b, e, h, k) and anti-talin antibody (in red, c, f, i, l). See Fig. 2 legend for panel layout.

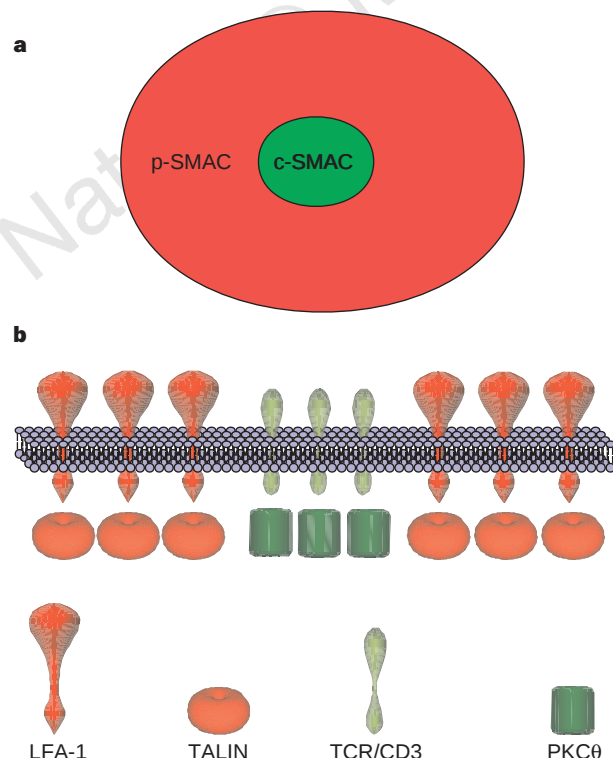


Figure 6 The p-SMAC and c-SMAC. **a**, 3D view of the entire cell contact. **b**, Cross-section through the centre of a T-cell-APC contact, showing the molecular clustering and segregation of the TCR, LFA-1, PKC- θ and talin.

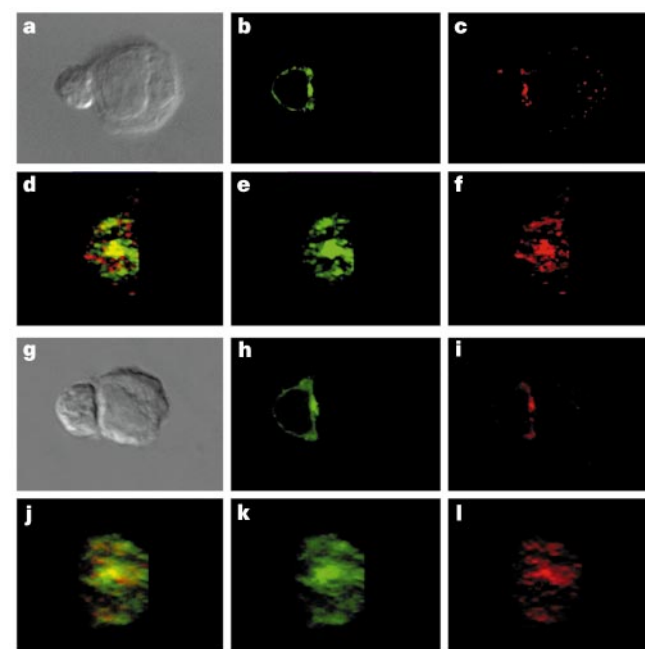


Figure 8 The rapid translocation of Fyn and Lck to c-SMACs. AD10-CH12 cell conjugates were fixed 5 (for Lck) or 13 (for Fyn) min after cell conjugation and labelled with anti-CD3 (in green, b, e, h, k), anti-Fyn (in red, c, f) or anti-Lck (in red, i, l) antibodies. See Fig. 2 legend for panel layout.

functionally uncoupled from the induction of T-cell proliferation^{3,8}. Importantly, in these conjugates the clustered CD3 and talin failed to segregate into SMACs and remained unorganized throughout the contact (Fig. 7j–l). The SMACs failed to form even when the CH12 cells were pulsed with higher concentrations of E8A (up to 50 µg ml⁻¹). These results indicate that receptor engagement by itself, although necessary, is not sufficient to form the SMACs and that receptor ligation without the formation of SMACs is not sufficient to properly activate the T cells.

The formation of SMACs may explain several important issues related to the physiological activation of T cells. The TCR–CD3 complexes do not remain randomly dispersed after binding ligand. Instead, they immediately begin to assemble into a c-SMAC (Fig. 8). Three-dimensional analysis indicated that, as early as 5–13 min after cell conjugation, the Src-family kinases Lck and Fyn were also enriched in the c-SMAC (Fig. 8). Other signalling proteins at the p-SMAC (A.K. *et al.*, manuscript in preparation) indicate that the p-SMAC may also act in T-cell signalling, albeit providing signals that differ from the c-SMAC signals. It is thus likely that the regulated segregation of activated receptors and their associated proteins would generate unique signals that would not be otherwise generated by randomly co-aggregated receptors.

Activation of receptor-associated tyrosine kinases and the appropriate tyrosine phosphorylation of the cytoplasmic tails of receptors and of multiple intracellular proteins^{1,2} requires the formation of receptor signalling complexes^{14–16}. The SMACs identify a new, additional level of regulated molecular interactions, whereby individual receptor signalling complexes appear to interact with each other to form distinct supramolecular complexes. Although the mechanisms that generate the SMACs are still unknown, it is likely that the SMACs are assembled by receptor interactions with extracellular ligands and with intracellular signalling and linking proteins, which can express multiple protein-interaction domains^{14–16}. We are now studying the mechanisms that regulate the formation of the SMACs and their roles in determining the fate of the activated T cells. □

Methods

The T and B cells, the formation of cell conjugates, the antibodies and the immunofluorescence labelling have been described^{3,7}. The agonist peptide CA37 and the altered peptides E8A and E8T, containing a substitution of the glutamic acid at position 8 with alanine or threonine, as well as all the D10-related cells, have been described¹³ and were a gift from C. A. Janeway. The anti-Iα antibody was a gift from J. Miller. P. Marrack provided the AD10–TCR transgenic mice. The three-dimensional immunofluorescence and the corresponding Nomarski images of the cells were recorded by a digital fluorescence microscopy system (Intelligent Imaging Innovations, Denver, CO). The imaging system included a chilled charge-coupled device (CCD) slow-scan camera (Orbis, SpectraSource, Westlake Village, CA), a Zeiss Axiophot microscope, a z-axis stepper motor (Della, Estes Park, CO), and excitation and emission filter wheels (LEP, Hawthorne, NY) with narrow-band optical filters (Chromatech, Brattleboro, VT). All of these components were controlled by SlideBook software (Intelligent Imaging Innovations). SlideBook software was used for three-dimensional image capture, deconvolution and rendering. At least 100 cell conjugates per coverslip were analysed. All the three-dimensional image sets were analysed using Nearest Neighbour Deconvolution. In addition, at least 25 of the same conjugates were also processed by the much slower Constrained Iterative Deconvolution, applying 20 iterations on the three-dimensional image sets. Additional iterations had no detectable effects on the images. There were no obvious differences between the images that were generated with the two different deconvolutions. The three-dimensional front views of the cell contacts were generated as orthographic projections of the voxels that make up the T-cell–APC contact.

Received 23 June; accepted 4 August 1998.

1. Weiss, A. & Littman, D. R. Signal transduction by lymphocyte antigen receptors. *Cell* **76**, 263–274 (1994).
2. Qian, D. & Weiss, A. T cell antigen receptor signal transduction. *Curr. Opin. Cell Biol.* **9**, 205–212 (1997).

3. Monks, C. R., Kupfer, H., Tamir, I., Barlow, A. & Kupfer, A. Selective modulation of protein kinase C- θ during T-cell activation. *Nature* **385**, 83–86 (1997).
4. Podack, E. R. & Kupfer, A. T-cell effector functions: mechanisms for delivery of cytotoxicity and help. *Annu. Rev. Cell Biol.* **7**, 479–504 (1991).
5. Singer, S. J. Intercellular communication and cell-cell adhesion. *Science* **255**, 1671–1677 (1992).
6. Kupfer, A. & Singer, S. J. Cell biology of cytotoxic and helper T cell functions: immunofluorescence microscopic studies of single cells and cell couples. *Annu. Rev. Immunol.* **7**, 309–337 (1989).
7. Hiraoka, Y., Sedat, J. W. & Agard, D. A. Determination of three-dimensional imaging properties of a light microscope system. Partial confocal behavior in epifluorescence microscopy. *Biophys. J.* **57**, 325–333 (1990).
8. Kupfer, A. & Singer, S. J. The specific interaction of helper T cells and antigen-presenting B cells. IV. Membrane and cytoskeletal reorganizations in the bound T cell as a function of antigen dose. *J. Exp. Med.* **170**, 1697–1713 (1989).
9. Kupfer, A., Burn, P. & Singer, S. J. The PMA-induced specific association of LFA-1 and talin in intact cloned T helper cells. *J. Mol. Cell. Immunol.* **4**, 317–325 (1990).
10. McCloskey, M. A. & Poo, M. M. Contact-induced redistribution of specific membrane components: local accumulation and development of adhesion. *J. Cell Biol.* **102**, 2185–2196 (1986).
11. Harding, C. V. & Unanue, E. R. Quantitation of antigen-presenting cell MHC class II/peptide complexes necessary for T-cell stimulation. *Nature* **346**, 574–576 (1990).
12. Anderson, M. S., Swier, K., Arneson, L. & Miller, J. Enhanced antigen presentation in the absence of the invariant chain endosomal localization signal. *J. Exp. Med.* **178**, 1959–1969 (1993).
13. Dittel, B. N., Sant'Angelo, D. B. & Janeway, C. A. Jr Peptide antagonists inhibit proliferation and the production of IL-4 and/or IFN- γ in T helper 1, T helper 2, and T helper 0 clones bearing the same TCR. *J. Immunol.* **158**, 4065–4073 (1997).
14. Schlessinger, J. SH2/SH3 signaling proteins. *Curr. Opin. Genet. Dev.* **4**, 25–30 (1994).
15. Lemmon, M. A., Ferguson, K. M. & Schlessinger, J. PH domains: diverse sequences with a common fold recruit signaling molecules to the cell surface. *Cell* **85**, 621–624 (1996).
16. Pawson, T. & Scott, J. D. Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**, 2075–2080 (1997).
17. Kupfer, H., Monks, C. R. & Kupfer, A. Small splenic B cells that bind to antigen-specific T helper (Th) cells and face the site of cytokine production in the Th cells selectively proliferate: immunofluorescence microscopic studies of Th-B-antigen presenting cell interactions. *J. Exp. Med.* **179**, 1507–1515 (1994).

Acknowledgements. We thank C. A. Janeway Jr, J. Miller and P. Marrack for reagents, cell lines and transgenic mice; J. Kappler and P. Marrack for comments; and D. Littman and J. Schlessinger for advice and critical reading of this manuscript. These studies were supported in part by grants from the NIH.

Correspondence and requests for materials should be addressed to A.K.

Yeast G1 cyclins are unstable in G1 phase

Brandt L. Schneider*†, E. Elizabeth Patton*‡, Stefan Lanker§||, Michael D. Mendenhall§, Curt Wittenberg§, Bruce Futcher† & Mike Tyers‡

† Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, New York 11724, USA

‡ Programme in Molecular Biology and Cancer, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Canada M5G 1X5, and Graduate Department of Molecular and Medical Genetics, University of Toronto, 1 Kings College Circle, Toronto, Canada M5S 1A8
§ Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

¶ Markey Cancer Center, Dept. of Biochemistry, University of Kentucky, Lexington, KY 40536-0096, USA

* These authors contributed equally to this work.

In most eukaryotes, commitment to cell division occurs in late G1 phase at an event called Start in the yeast *Saccharomyces cerevisiae*¹, and called the restriction point in mammalian cells². Start is triggered by the cyclin-dependent kinase Cdc28 and three rate-limiting activators, the G1 cyclins Cln1, Cln2 and Cln3 (ref. 3). Cyclin accumulation in G1 is driven in part by the cell-cycle-regulated transcription of *CLN1* and *CLN2*, which peaks at Start³. *CLN* transcription is modulated by physiological signals that regulate G1 progression^{4,5}, but it is unclear whether Cln protein stability is cell-cycle-regulated. It has been suggested that once cells pass Start, Cln proteolysis is triggered by the mitotic cyclins Clb1, 2, 3 and 4 (ref. 6). But here we show that G1 cyclins are unstable in G1 phase, and that Clb–Cdc28 activity is not needed for G1 cyclin turnover. Cln instability thus provides a means to couple Cln–Cdc28 activity to transcriptional regulation and protein synthetic rate in pre-Start G1 cells.

|| Present address: Department of Molecular and Medical Genetics, School of Medicine, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97201-3098, USA.

Cln1, Cln2 and Cln3 are non-abundant proteins that efficiently promote the G1/S transition. Direct measurement of Cln half-lives in G1 phase has been difficult. The Clns are very unstable proteins with half-lives of less than 10 min in asynchronous populations⁷⁻¹⁴, consistent with the idea that Cln instability is constitutive. Indeed, Clns are unstable in G1 cells arrested with mating pheromone^{9,15,16}, but this could be a special effect of pheromone rather than being representative of G1 cells in logarithmic growth. However, it is possible that Cln1 and Cln2 are stable in late G1 and are degraded shortly after cells pass through Start.

To examine the G1-phase stability of Cln2 directly, we measured the half-life of Cln2 in small G1-phase cells obtained by elutriation. We used strains that contained either *CLN2*^{HA} (which is phenotypically indistinguishable from untagged *CLN2* (see Fig. 1b)) or *CLN2*^{4T3S-HA}, both expressed from the *GAL1* promoter. *CLN2*^{4T3S} is a multiple point mutant that lacks seven consensus Cdc28 phosphorylation sites and is sevenfold more stable than wild-type Cln2 in asynchronous cultures¹². The half-life of the mutant and wild-type proteins were measured in G1 cells by decay of the Cln2 signal after repression of the *GAL1* promoter by glucose. Wild-type Cln2 had a half-life of 5–10 min in the G1 cells (Fig. 1a), which remained in G1 phase throughout the experiment. In contrast, the stabilized mutant *CLN2*^{4T3S} had a longer half-life, and a fraction of the cells budded and began DNA synthesis (Fig. 1a). This experiment shows

directly that Cln2 is unstable in G1 phase and suggests that Cln2 instability is required for proper regulation of Start.

One caveat to this experiment is that the instability of Cln2 might have been due partly to the HA epitope tag, which was required to detect the low levels of wild-type Cln2 in small G1 cells. However, we compared the stability of HA-tagged versus untagged Cln2 and found that the half-lives of these two forms of Cln2 were indistinguishable in asynchronous cultures (Fig. 1b). We also determined that neither tag position nor the epitope sequence altered Cln2 stability (not shown). Furthermore, we compared the ability of untagged *GAL1-CLN2* and untagged *GAL1-CLN2*^{4T3S} to promote Start. Small G1 cells were obtained by elutriation, and then the *GAL1* promoter was turned on slightly with a low concentration of galactose in a semi-repressing medium. G1 expression of Cln2^{4T3S} accelerated Start and reduced critical cell size compared to the expression of wild-type Cln2 (Fig. 1c). The differential ability of stabilized Cln2 to promote Start again suggests that wild-type Cln2 is unstable in G1 cells.

It has recently been suggested that the Clb mitotic cyclins are necessary for Cln degradation⁶, a model that predicts that G1 cyclins should be stable in pre-Start G1 cells because such cells lack Clb protein entirely¹⁷. Because this prediction is inconsistent with our finding that Cln2 is unstable in G1 cells, we measured Cln2 stability under three different conditions that severely reduce or eliminate Clb–Cdc28 activity. Cells lacking Clb1 to Clb4 (strains MT383 and K3390, which we call '*clb1-4*^{ts}')¹⁸ arrest in G2 phase, whereas cells lacking Clb1 to Clb6 (strain K4057, which we call '*clb1-6*^{ts}')¹⁹ arrest in G1 phase. Cells that overexpress the Clb–Cdc28 inhibitor Sic1 also accumulate in G1 phase^{20,21}. When Cln stability was measured in a *clb1-4*^{ts} strain by repression of a *GAL1-CLN2*^{HA} construct (Fig. 2a), or a *GAL1-CLN1*^{HA} construct (not shown), Cln was no

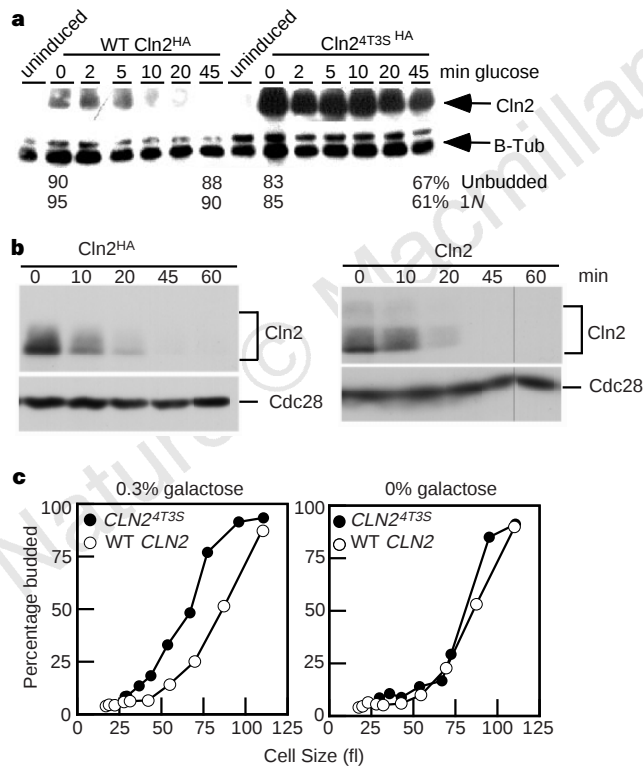


Figure 1 Cln2 is unstable in G1 phase. **a**, Direct measurement of Cln2 stability in G1 cells. The decay of wild-type (WT) *CLN2*^{HA} or the stable mutant *CLN2*^{4T3S} in synchronized G1-phase cells after repression of weakly induced *GAL1-CLN2* by glucose was measured by immunoblotting against the HA epitope. Cells in G1 phase were assayed as the percentage of unbudded cells, and as the percentage of cells with 1N DNA content. B-tubulin immunoreactivity was used as a loading control. **b**, The HA epitope tag does not affect the stability of Cln2. Tagged Cln2 and untagged Cln2 were expressed from the *GAL1* promoter in asynchronous cultures. Cln2 decay after repression of the *GAL1* constructs by 2% glucose was assessed by immunoblotting with an anti-Cln2 rabbit polyclonal antibody. **c**, *CLN2*^{4T3S} (untagged) is a better inducer of Start than wild-type *CLN2* (untagged). *GAL1-CLN2*^{4T3S} or *GAL1-CLN2* was weakly induced in small G1 cells and the rate of passage through Start was assessed by bud formation as a function of cell size.

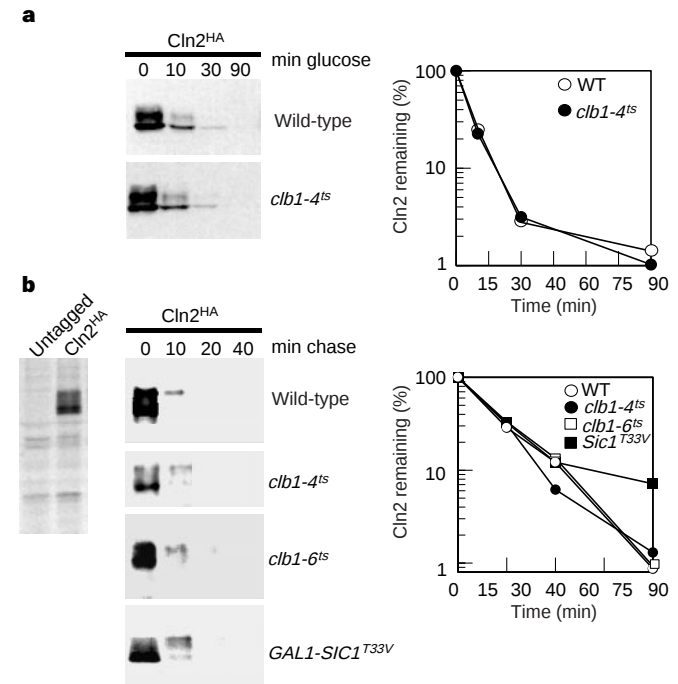


Figure 2 Clb activity is not required for Cln2 degradation. **a**, Cln2 is unstable in a *clb1-4*^{ts} strain. *GAL1-CLN2*^{HA} was expressed in a wild-type (WT) or *clb1-4*^{ts} strain at a non-permissive temperature, and repressed by glucose for the indicated times. *CLN2*^{HA} immunoreactivity was quantified by densitometry and normalized to Cdc28 immunoreactivity on the same blot. **b**, Cln2 is unstable under three different conditions in which Clb activity is severely compromised. *CLN2*^{HA} was expressed at wild-type levels from the *CLN2* promoter in *clb1-4*^{ts}, *clb1-6*^{ts} and *GAL1-SIC1*^{T33V} strains under non-permissive conditions. Stability of Cln2^{HA} was determined by [³⁵S]methionine/cysteine pulse-chase analysis.

more stable than in wild-type cells. To rule out possible effects of *CLN* overexpression and carbon source shifts on Cln stability, we performed pulse-chase analysis of ^{35}S -labelled Cln2 expressed at wild-type levels from its own promoter. Cln2 was as unstable in the *clb1-4^{ts}* and *clb1-6^{ts}* arrested cells as in the wild-type strain (Fig. 2b). Cln2 was also unstable in cells overexpressing *SIC1^{T33V}*, which encodes a partly stabilized version of Sic1 (Fig. 2b). Furthermore, we failed to detect any alteration in Cln1 half-life in cells arrested by overexpression of *SIC1^{T45A}* (not shown), which encodes a highly stable version of Sic1. Therefore, the mitotic cyclins cannot be essential for G1 cyclin degradation.

Finally, we re-examined the role of the ubiquitin-conjugating enzyme Cdc34 in the turnover of Cln2. There is persuasive evidence that Cdc34 has a direct role in conjugating ubiquitin to Cln2 (refs. 11, 14) and to Sic1 (refs 20, 22, 23), leading to degradation. For ubiquitination of Sic1, Cdc34 acts in concert with an E3 ubiquitin protein ligase complex composed of Skp1, Cdc53 and the F-box protein Cdc4 (the SCF^{Cdc4} complex)^{22,23}. Cln ubiquitination proceeds via a similar pathway, but depends on a different F-box protein, Grr1, which forms the analogous SCF^{Grr1} complex^{13,22,24}. However, it has been argued that Cdc34 promotes Cln degradation indirectly by promoting the degradation of Sic1, thereby activating Clb kinases⁶. If this were true, then *cdc4* mutants should also be defective in Cln2 turnover, because Sic1 accumulates in *cdc4* mutants just as it does in *cdc34* mutants¹⁹. However, Cln2 degradation was defective only in the *cdc34* mutant (Fig. 3a), despite the fact that Sic1 accumulated in both the *cdc34* mutant and the *cdc4* mutant (Fig. 3b). The fact that the *cdc34* arrest phenotype is indistinguishable from *cdc4* and *clb1-6^{ts}* arrests, yet Cln2 is stabilized only in the *cdc34* arrest, indicates that Cln stabilization in *cdc34* mutants is not an indirect effect of Sic1 accumulation and loss of Clb activity.

A model in which G1 cyclins are stable in G1 phase and then catastrophically destroyed as cells pass Start, is attractive because it is analogous to the cell-cycle-regulated destruction of mitotic cyclins in anaphase¹⁷. However, our data rule out this model and support a model in which Cln proteins are constitutively unstable. On the basis of our data we suggest that the rise in Cln abundance is due to a rise in the rate of Cln synthesis, which depends both on a rise in *CLN* transcription and on increases in overall protein synthesis as cells grow. After Start, the onset of mitotic cyclin activity represses Cln synthesis, but only at the transcriptional level¹⁸. The instability of the Cln proteins probably derives from Cdc28-dependent autophosphorylation of the Cln subunit, which targets the Cln for degradation^{10-12,14}. The intrinsic instability of the Cln proteins is advantageous for the cell because it allows almost instantaneous responses to changes in the rate of protein synthesis²⁵ and *CLN* transcription^{4,5}, both of which are key determinants of

whether or not cell division is appropriate. The continuous requirements for growth factors and protein synthesis in G1-phase progression of mammalian cells²⁶, and the phosphorylation-dependent instability of cyclin D and cyclin E (refs 27-29), suggest that a similar control of G1 cyclin activity operates in mammalian cells. □

Methods

Cell synchronization. All strains were isogenic with W303 (*adc2 his3 leu2 trp1 URA3*). To determine the half-life of Cln2 in G1 cells, strain BS328 (*cln3*) containing either plasmid pSL46 (ref. 12) (*GAL1-CLN2^{HA} URA3 CEN*, strain BS333) or plasmid pSL122 (ref. 12) (*GAL1-CLN2^{4T3S-HA} URA3 CEN*, strain BS322) was grown to early log phase in yeast extract peptone medium plus 2% sucrose. The use of a *cln3* deletion strain, which has a prolonged G1 phase¹⁵, was required to allow sufficient time for half-life time courses to be completed before cells passed Start. Small G1 cells were obtained by elutriation at room temperature in clarified medium. The *GAL1* promoter was weakly induced for 1 h by the addition of galactose in the presence of 2% sucrose (a semi-repressing condition) and then repressed by the addition of 2% glucose. The decay of Cln2^{HA} or Cln2^{4T3S-HA} was assayed by immunoblotting as described¹⁵.

To assess the ability of Cln2 to activate Start, strain BS328 containing plasmid pB75 (*GAL1-CLN2 URA3 CEN*, strain BS412) or plasmid pB101 (*GAL1-CLN2^{4T3S} URA3 CEN*, strain BS410) were grown and elutriated as above. After G1 cells were obtained, galactose (0.3%) was either added or not added in the presence of 2% sucrose. Passage through Start was assessed by monitoring bud formation by microscopy, and modal cell size with a Coulter Channelyzer¹⁵.

Protein analysis. For determination of Cln2 half-life by promoter shut-off, strain MT386 (*cln2::GAL1-CLN2^{HA}-LEU2*) and an isogenic *clb1-4^{ts}* strain MT383 (*clb1 clb2-VI clb3::TRP1 clb4::HIS3 cln2::GAL1-CLN2^{HA}-LEU2*)¹⁸ were grown in 2% raffinose, shifted to 37 °C for 2 h, induced with 2% galactose for 1.5 h at 37 °C, then repressed with 2% glucose at 37 °C for the indicated times. Similar results were also obtained at a lower non-permissive temperature of 34 °C. Cln2 decay was assayed by immunoblotting with 12CA5 antibody using detection by enhanced chemoluminescence as described¹⁵. Polyclonal antibodies directed against Sic1, Cln2, Cdc28 and B-tubulin were used as described^{15,22,30}.

For determination of Cln2 half-life by ^{35}S pulse-chase analysis, all strains contained a single copy of *CLN2^{HA}* expressed from the *CLN2* promoter at wild-type levels. Strains used were: K3389 (*CLN2^{HA}*)¹⁸, K3390 (*clb1 clb2-VI clb3 clb4 CLN2^{HA}*)¹⁸, K4057 (*clb1 clb2-VI clb3::TRP1 clb4::HIS3 clb5::GAL-CLB5-URA3 clb6::LEU2*)¹⁹, MT668 (*cdc4-1*) and MT670 (*cdc34-2*). K4057 was transformed with plasmid pMT1634 (*CLN2^{HA} ADE2 CEN*); K3389 was transformed with plasmid pMDM202 (*GAL1-SIC1^{T33V} LEU2 CEN*); W303, MT668 and MT670 were transformed with plasmid pMT291 (*CLN2^{HA} LEU2 CEN*). For experiments with temperature-sensitive strains, cultures were shifted to 37 °C for 2 h before the pulse-chase regimen. Strain K4057, which requires induction of *GAL1-CLB5* with 0.3% galactose for viability¹⁹, was simultaneously shifted to glucose medium at the time of the shift to 37 °C. *CLN2^{HA}* stability in strain K3389 was measured at 30 °C after induction of *GAL1-SIC1^{T33V}* with 2% galactose for 2 h.

The pulse-chase regimen, protein extraction and immunoprecipitation were based on ref. 17. In brief, 15 ml of culture at 0.5×10^7 cells ml⁻¹ in rich medium supplemented with 0.1 mM methionine was filtered and resuspended in 1 ml of -Met synthetic medium. After 5 min, cultures were pulsed with 500 μCi of [^{35}S]Met/Cys (ICN, Tran ^{35}S -label) and after an additional 5 min cultures were chased with 2 mM unlabelled Met, 2 mM unlabelled Cys, and 1 mg ml⁻¹ cycloheximide. Cells were pelleted, resuspended in 5% cold trichloroacetic acid (TCA), washed once with cold acetone and stored at -80 °C. Frozen pellets were resuspended in 75 μl of cold breaking buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, and protease inhibitors: 1 mM PMSF, 0.6 mM dimethylaminopurine, 1 μg ml⁻¹ leupeptin, 1 μg ml⁻¹ pepstatin, 10 μg ml⁻¹ TPCK, 10 μg ml⁻¹ soybean trypsin inhibitor) with glass beads and broken by vortex mixing, after which SDS was added to 1% and the lysates boiled for 2 min. Lysates were diluted to 1 ml with 900 μl RIPA buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 150 mM NaCl, 0.5% sodium deoxycholate, 1% NP-40 and protease inhibitors) plus 0.2 mg ml⁻¹ BSA and cell debris was pelleted. We determined that the above lysis buffers were critical to maintain the Cln phosphoisoforms when compared with other lysis buffers⁶. Incorporation of

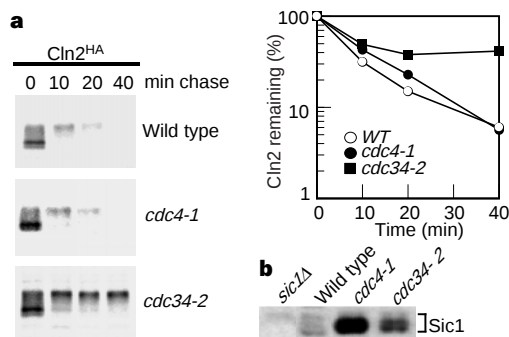


Figure 3 Sic1 accumulation does not affect Cln2 stability. **a**, Cdc34, but not Cdc4, function is required for Cln2 degradation. *CLN2^{HA}* was expressed at wild-type levels from the *CLN2* promoter in wild-type, *cdc4-1* and *cdc34-2* strains at the non-permissive temperature. Stability of Cln2^{HA} was determined by [^{35}S]methionine/cysteine pulse-chase analysis. **b**, Abundance of Sic1 was determined in lysates from each strain in **a** by immunoblotting with anti-Sic1 antibody.

³⁵S was measured by precipitation with TCA, and equal radioactive counts were used for each immunoprecipitation. The supernatant was precleared with 20 µl protein A–Sepharose beads for 20 min, incubated with 1 µg 12CA5 anti-HA antibody for 1 h and immune complexes collected on 10 µl protein A–Sepharose beads for 2 h. Beads were washed twice with 1 ml RIPA buffer, twice with 1 ml RIPA buffer plus 1% β-mercaptoethanol, and twice with 1 ml RIPA buffer plus 1% β-mercaptoethanol plus 2 M urea. Beads were transferred to a fresh tube after every second wash. Beads were aspirated dry, resuspended in 10 µl cold 2× Laemmli sample buffer, boiled for 2 min and proteins were separated on a 10% SDS–PAGE gel. Radioactive species were visualized and quantitated using a PhosphorImager (Molecular Dynamics). Incorporation into Cln2-specific species was normalized to total radioactivity in parallel lanes of total lysate.

Received 16 January; accepted 2 July 1998.

- Hartwell, L. H., Culotti, J., Pringle, J. R. & Reid, B. J. Genetic control of the cell division cycle in yeast. *Science* **183**, 46–51 (1974).
- Rosow, P. W., Riddle, V. G. & Pardee, A. B. Synthesis of labile, serum-dependent protein in early G1 controls animal cell growth. *Proc. Natl Acad. Sci. USA* **76**, 4446–4450 (1979).
- Cross, F. R. Starting the cell cycle: what's the point? *Curr. Opin. Cell Biol.* **7**, 790–797 (1995).
- Baroni, M. D., Monti, P. & Alberghina, L. Repression of growth-regulated G1 cyclin expression by cyclic AMP in budding yeast. *Nature* **371**, 339–342 (1994).
- Tokiwa, G., Tyers, M., Volpe, T. & Futcher, B. Inhibition of G1 cyclin activity by the Ras/cAMP pathway in yeast. *Nature* **371**, 342–345 (1994).
- Blondel, M. & Mann, C. G2 cyclins are required for the degradation of G1 cyclins in yeast. *Nature* **384**, 279–282 (1996).
- Tyers, M., Tokiwa, G., Nash, R. & Futcher, B. The Cln3–Cdc28 kinase complex of *S. cerevisiae* is regulated by proteolysis and phosphorylation. *EMBO J.* **11**, 1773–1784 (1992).
- Cross, F. R. & Blake, C. M. The yeast Cln3 protein is an unstable activator of Cdc28. *Mol. Cell. Biol.* **13**, 3266–3271 (1993).
- Salama, S. R., Hendricks, K. B. & Thorner, J. G1 cyclin degradation: the PEST motif of yeast Cln2 is necessary, but not sufficient, for rapid protein turnover. *Mol. Cell. Biol.* **14**, 7953–7966 (1994).
- Yaglom, J. *et al.* p34^{Cdc28}-mediated control of Cln3 cyclin degradation. *Mol. Cell. Biol.* **15**, 731–741 (1995).
- Deshais, R. J., Chau, V. & Kirschner, M. Ubiquitination of the G1 cyclin Cln2p by a Cdc34p-dependent pathway. *EMBO J.* **14**, 303–312 (1995).
- Lanker, S., Valdivieso, M. H. & Wittenberg, C. Rapid degradation of the G1 cyclin Cln2 induced by CDK-dependent phosphorylation. *Science* **271**, 1597–1601 (1996).
- Barral, Y., Jentsch, S. & Mann, C. G1 cyclin turnover and nutrient uptake are controlled by a common pathway in yeast. *Genes Dev.* **9**, 399–409 (1995).
- Willems, A. R. *et al.* Cdc53 targets phosphorylated G1 cyclins for degradation by the ubiquitin proteolytic pathway. *Cell* **86**, 453–463 (1996).
- Tyers, M. & Futcher, B. Far1 and Fus3 link the mating pheromone signal transduction pathway to three G1-phase Cdc28 kinase complexes. *Mol. Cell. Biol.* **13**, 5659–5669 (1993).
- Valdivieso, M. H., Sugimoto, K., Jahng, K. Y., Fernandes, P. M. & Wittenberg, C. FAR1 is required for posttranscriptional regulation of CLN2 gene expression in response to mating pheromone. *Mol. Cell. Biol.* **13**, 1013–1022 (1993).
- Amon, A., Imiger, S. & Nasmyth, K. Closing the cell cycle circle in yeast: G2 cyclin proteolysis initiated at mitosis persists until the activation of G1 cyclins in the next cycle. *Cell* **77**, 1037–1050 (1994).
- Amon, A., Tyers, M., Futcher, B. & Nasmyth, K. Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins. *Cell* **74**, 993–1007 (1993).
- Schwob, E., Bohm, T., Mendenhall, M. D. & Nasmyth, K. The B-type cyclin kinase inhibitor p40^{SCC1} controls the G1 to S transition in *S. cerevisiae*. *Cell* **79**, 233–244 (1994).
- Verma, R. *et al.* Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase. *Science* **278**, 455–460 (1997).
- Nugroho, T. T. & Mendenhall, M. D. An inhibitor of yeast cyclin-dependent protein kinase plays an important role in ensuring the genomic integrity of daughter cells. *Mol. Cell. Biol.* **14**, 3320–3328 (1994).
- Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J. & Harper, J. W. F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin–ligase complex. *Cell* **91**, 209–219 (1997).
- Feldman, R. M., Correll, C. C., Kaplan, K. B., Deshaies, R. J. A complex of Cdc4p, Skp1p, and Cdc53p/cul1 catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. *Cell* **91**, 221–230 (1997).
- Patton, E. E. *et al.* Cdc53 is a scaffold protein for multiple Cdc34/Skp1/F-box protein complexes that regulate cell division and methionine biosynthesis in yeast. *Genes Dev.* **12**, 692–705 (1998).
- Ko, H. A. & Moore, S. A. Kinetic characterization of a prestart cell division control step in yeast. Implications for the mechanism of alpha-factor-induced division arrest. *J. Biol. Chem.* **265**, 21652–21663 (1990).
- Zetterberg, A. & Larsson, O. Coordination between cell growth and cell cycle transit in animal cells. *Cold Spring Harb. Symp. Quant. Biol.* **56**, 137–147 (1991).
- Clurman, B. E., Sheaff, R. J., Thress, K., Groudine, M. & Roberts, J. M. Turnover of cyclin E by the ubiquitin–proteasome pathway is regulated by cdk2 binding and cyclin phosphorylation. *Genes Dev.* **10**, 1979–1990 (1996).
- Diehl, J. A., Zindy, F. & Sherr, C. J. Inhibition of cyclin D1 phosphorylation on threonine-286 prevents its rapid degradation via the ubiquitin–proteasome pathway. *Genes Dev.* **11**, 957–972 (1997).
- Wong, K. A. & Reed, S. I. Activation of cyclin E/CDK2 is coupled to site-specific autophosphorylation and ubiquitin-dependent degradation of cyclin E. *EMBO J.* **15**, 4182–4193 (1996).
- Wittenberg, C., Sugimoto, K. & Reed, S. I. G1-specific cyclins of *S. cerevisiae*: cell cycle periodicity, regulation by mating pheromone, and association with the p34CDC28 protein kinase. *Cell* **62**, 225–237 (1990).

Acknowledgements. We thank A. Amon for advice on pulse–chase experiments, E. Schwob and K. Nasmyth for strains, and C. Mann, M. Blondel and D. Kornitz for sharing unpublished data and many helpful discussions. B.S. was supported by an Army Breast Cancer post-doctoral fellowship, E.P. by a National Sciences and Engineering Research Council pre-doctoral fellowship, and S.L. by the Swiss National Science Foundation and the Human Frontiers in Science program. M.D.M., C.W. and B.F. are supported by grants from the NIH, and M.T. is supported by grants from the Medical Research Council of Canada and the National Cancer Institute of Canada (NCIC) and is a Research Scientist of the NCIC.

Correspondence and requests for materials should be directed to B.F. (email: futcher@cshl.org) and M.T. (e-mail: tyers@mshri.on.ca).

DNA hypomethylation leads to elevated mutation rates

Richard Z. Chen*, Ulf Pettersson*†, Caroline Beard*†, Laurie Jackson-Grusby* & Rudolf Jaenisch*‡

* Whitehead Institute for Biomedical Research and ‡ Massachusetts Institute of Technology, Department of Biology, Cambridge, Massachusetts 02142, USA

Genome-wide demethylation has been suggested to be a step in carcinogenesis¹. Evidence for this notion comes from the frequently observed global DNA hypomethylation in tumour cells², and from a recent study suggesting that defects in DNA methylation might contribute to the genomic instability of some colorectal tumour cell lines³. DNA hypomethylation has also been associated with abnormal chromosomal structures, as observed in cells from patients with ICF (Immunodeficiency, Centromeric instability and Familial abnormalities) syndrome^{4,5} and in cells treated with the demethylating agent 5-azadeoxycytidine⁶. Here we report that murine embryonic stem cells nullizygous for the major DNA methyltransferase (*Dnmt1*) gene exhibited significantly elevated mutation rates at both the endogenous hypoxanthine phosphoribosyltransferase (*Hprt*) gene and an integrated viral thymidine kinase (*tk*) transgene. Gene deletions were the predominant mutations at both loci. The major cause of the observed *tk* deletions was either mitotic recombination or chromosomal loss accompanied by duplication of the remaining chromosome. Our results imply an important role for mammalian DNA methylation in maintaining genome stability.

Disruption of the *Dnmt1* gene in mice results in substantial genomic demethylation and an embryonic lethal phenotype^{7,8}. Significantly, *Dnmt1*^{−/−} embryonic stem (ES) cells proliferate normally despite their DNA being highly demethylated (refs 7, 8; and data not shown). This allowed us to perform genetically controlled experiments to investigate the effects of DNA hypomethylation on genome stability. We therefore compared the spontaneous mutation rates at two selectable genes, the endogenous X-linked *Hprt* gene and an autosomal viral *tk* transgene, in isogenic male ES cell lines either proficient or deficient in DNA methylation by screening large ES cell populations for drug-resistant clones.

Hprt: *Hprt* mutants were identified by 6-thioguanine (6-TG) selection, which screens for mutations that impair the function of the enzyme. As summarized in Table 1a, significantly more 6-TG-resistant colonies were detected in two independently derived *Dnmt1*^{−/−} ES lines (S/S and C/C) than in two independently isolated wild-type ES cell lines (J1 and V18), or in a *Dnmt1*^{+/−} ES cell line (C/+), which showed a normal level of genomic methylation (ref. 8; and data not shown) and from which the homozygous mutant C/C line had been derived⁸. As control, we performed 6-TG selection on an ES cell line (C/CHIP) in which the *Dnmt1*^{−/−} genetic defect had been rescued by a *Dnmt1* cDNA (ref. 9). This 'cDNA-rescued' ES cell line exhibited the same low *Hprt* mutant frequency that was observed in the wild-type ES cell lines (Table 1a), indicating that disruption of *Dnmt1* was the genetic cause of the increased mutations in *Dnmt1*^{−/−} ES cells. To quantitate the effects of DNA hypomethylation on *Hprt* mutational rate, we applied the *P*₀ method of the Luria–Delbrück fluctuation analysis¹⁰, which allows the estimation of mutation rate from independent drug selections. By this method, DNA-methylation-proficient (J1, V18, C/+ and C/CHIP) or deficient (C/C) ES cells showed a mutation rate of 6.0×10^{-9} or of 6.3×10^{-8} , respectively (Table 1a). Thus,

† Present addresses: Department of Medical Genetics, University of Uppsala Biomedical Center, Box 589, S-75123 Uppsala, Sweden (U.P.); Ariad Pharmaceuticals, Inc., 26 Landsdowne Street, Cambridge, MA 02139, USA (C.B.).

the *Dnmt1*^{-/-}-induced genomic hypomethylation increases *Hprt* mutation rate by at least 10-fold.

We next used a combination of Southern blotting and reverse transcription-mediated polymerase chain reaction (RT-PCR) to study the nature of *Hprt* mutations in the 6-TG-resistant clones (representative data shown in Fig. 1). We analysed 28 independent 6-TG-resistant clones from *Dnmt1*^{-/-} (C/C) ES cells and 5 from methylation-proficient (J1, V18 or C/+) ES cells. Southern analysis revealed genomic rearrangements in a large fraction of the 6-TG-resistant C/C clones (15 of 28, as shown, for example, in lanes 2, 3, 4, 5, 6, 9, 10 and 11 in Fig. 1b). By RT-PCR, three additional 6-TG-resistant C/C clones (Fig. 1c, lanes 2, 8 and 17) showed alterations involving at least two exons, suggesting deletion of genomic regions that include these exons. Thus, the Southern and RT-PCR data suggest that the predominant pattern of *Hprt* mutations in DNA-

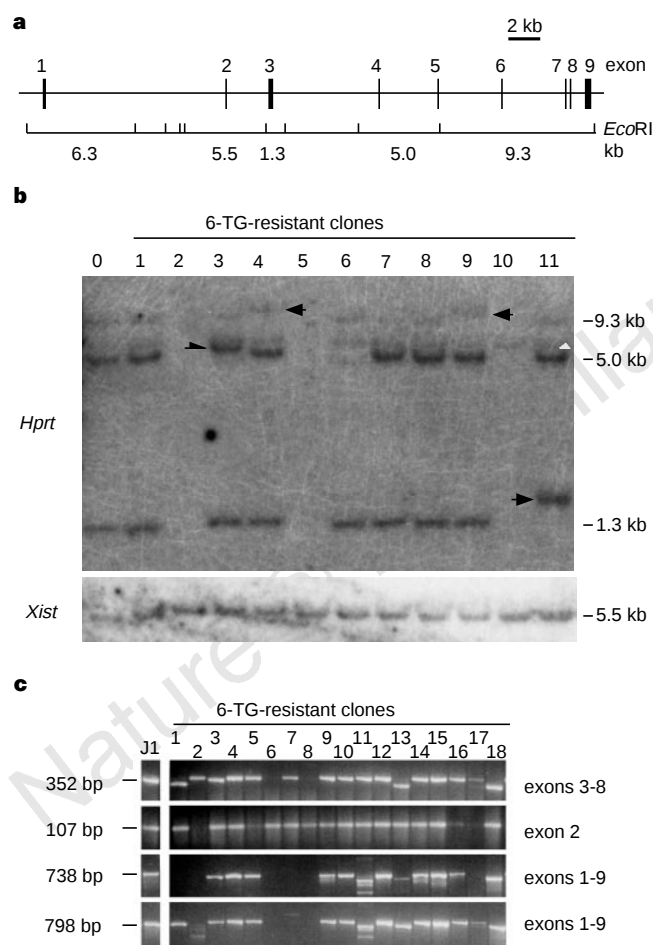


Figure 1 Analysis of *Hprt* mutations in the 6-TG-resistant ES clones. **a**, the *Hprt* genomic locus. Black boxes indicate exons drawn to approximate scale. *EcoRI* sites (small vertical lines) and the sizes of exon-containing *EcoRI* fragments are indicated. **b**, Southern analysis of *EcoRI*-digested genomic DNA with an *Hprt* cDNA probe (spanning exons 1-8). A total of 28 independent 6-TG-resistant clones from *Dnmt1*^{-/-} (C/C) and 5 from methylation-proficient (J1, V18 or C/+) ES cells were analysed. Shown are the results of an analysis with DNA samples from 11 6-TG-resistant C/C ES clones (lanes 1-11) and the control DNA from unselected C/C ES cells (lane 0). Hybridization of an *Xist* genomic probe to the blot confirms the presence of X-chromosomal DNA in each lane. Upshifted bands in lanes 3, 4, 9 and 11 are indicated by arrows. **c**, RT-PCR analysis. Shown are the results with RNA samples from 18 6-TG-resistant C/C ES clones (lanes 1-18) and the control RNA from wild-type J1 ES cells (lane J1). The order of the lanes does not necessarily correspond to that in **b**. Clones that showed genomic arrangements in the Southern analysis also demonstrated alterations in RT-PCR (note, for example, that lanes 3 and 6 in **c** correspond to lanes 4 and 9, respectively, in **b**).

methylation-deficient ES cells comprised deletions and insertions (at least 18 out of 28, or 64%). Because only five 6-TG-resistant clones from methylation-proficient ES cells were available for analysis (two showed deletions), it remains unclear whether hypomethylation causes a relative increase of genomic rearrangements at the *Hprt* locus. Our data do, however, indicate that the absolute frequency of rearrangements is increased by at least 10-fold.

tk: To investigate the effect of DNA hypomethylation on large-scale intra- and/or interchromosomal rearrangements, we measured the mutation rate of an autosomal gene. For this, a single copy of the herpes simplex virus *tk* gene was introduced through homologous recombination onto chromosome 9 of a 129/C57 F1 *Dnmt1*^{+/-} ES cell line. In this cell line, markers on chromosome 9 that are polymorphic between the two mouse strains would facilitate the analysis of rearrangements between the two homologous

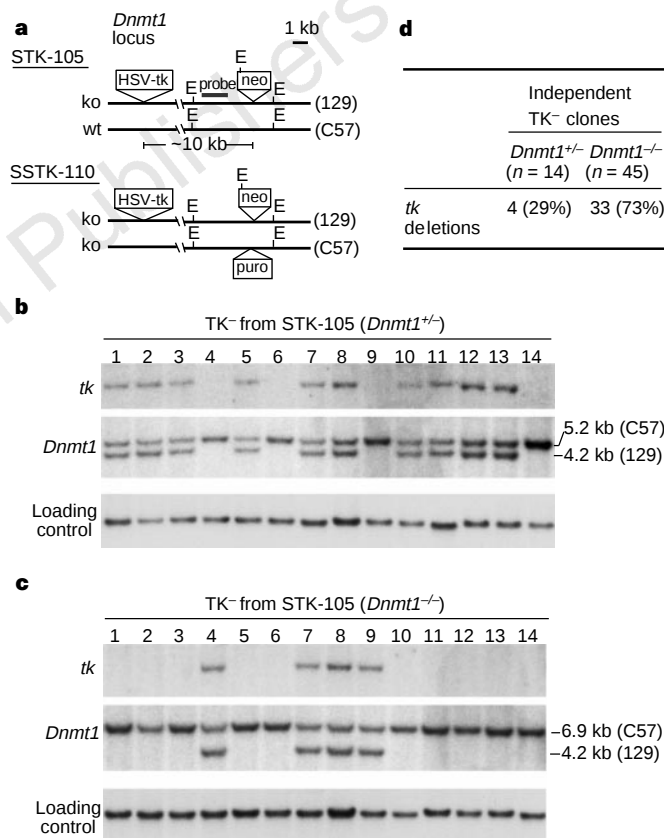


Figure 2 Creation of a hemizygous HSV-*tk* transgene locus and analysis of *tk* mutations in ganciclovir-resistant ES cells. **a**, An integrated HSV-*tk* transgene on either *Dnmt1*^{+/-} (STK-105) or *Dnmt1*^{-/-} (SSK-110) background. Shown is the *Dnmt1* locus where a single copy of HSV-*tk* was inserted (on the 129 chromosome). E, *EcoRI*. Southern analysis of *EcoRI*-digested genomic DNA with the indicated *Dnmt1* probe detects a 5.2-kb wild-type (wt) *Dnmt1* allele, a 4.2-kb neo-containing mutant (ko) *Dnmt1* allele and a 6.9-kb puro-containing mutant *Dnmt1* allele. **b**, **c**, *tk* gene deletions in ganciclovir-resistant (TK-) clones from STK-105 (**b**) or SSK-110 (**c**) ES cells. A Southern blot of *EcoRI*-digested genomic DNAs from 14 independent TK- clones of STK-105 and 45 of SSK-110 (results from 14 are shown) was hybridized with a *tk* probe (top panels) to detect *tk* gene deletions, with the *Dnmt1* probe (middle panels) to detect changes involving a linked genomic region, and with a *lacI* probe (bottom panels; both cell lines contain *lacI* transgenes elsewhere in their genomes²⁵) to control for the amount of DNA loaded. **d**, Summary of the frequencies of *tk* gene deletions in the TK- clones from STK-105 (*Dnmt1*^{+/-}) or SSK-110 (*Dnmt1*^{-/-}) cells. *n* is the total number of independent TK- clones analysed. The statistical significance of the difference was determined by χ^2 test ($P < 0.003$).

chromosomes. Insertion of the *tk* gene occurred at a site approximately 10 kb upstream of a *Dnmt1* mutation (a PGKneo insertion; Fig. 2a) on the 129 chromosome. A *tk*-containing *Dnmt1*^{-/-} ES cell line (SSTK-110) was generated by a second targeting of the wild-type *Dnmt1* allele on the C57 chromosome, and a non-targeted heterozygous clone (STK-105) from the transfection was used as a control cell line. Using a similar strategy to that in the Hprt assay, we subjected independently expanded populations of SSTK-110 and STK-105 cells to ganciclovir selection, which screens for loss of TK activity. By this assay, SSTK-110 cells demonstrated a significant increase (~6-fold; Table 1b) in mutation rate at the *tk* transgene compared with STK-105 control cells, which is consistent with a mutator phenotype of *Dnmt1*^{-/-} ES cells.

To study the mechanisms of *tk* gene inactivation, we determined the karyotypes of 12 representative *Dnmt1*^{-/-} TK mutants, 7 of which contained *tk* gene deletions. No aneuploidies were observed (not shown), suggesting that the *tk* mutations were unlikely to be caused by extensive chromosomal instability. We then analysed by Southern blotting 14 and 45 independent TK mutant clones from STK-105 and SSTK-110 cells, respectively. Most of the TK mutants (33 of 45; 73%) from the homozygous SSTK-110 cells had lost the *tk* gene completely, whereas only a minor portion of the TK mutants (4 of 14; 29%) from the heterozygous STK-105 cells lacked the *tk* gene (Fig. 2b, c, upper panels; summarized in Fig. 2d). All the *tk*-deleted clones, from either ES cell line, had also lost the linked 129 *Dnmt1* allele (Fig. 2b, c, middle panels). Furthermore, most *tk*-deleted clones from either cell line contained two copies of the C57 *Dnmt1* locus (Fig. 2b, c; compare band intensities in the middle panels using those of the lower panels as the loading control). These results suggest that most *tk* gene deletions were caused by genomic rearrangements involving the C57 chromosome that had led to loss of heterozygosity (LOH) of regions covering the *tk* gene but spanning unknown distances on the 129 chromosome.

To examine the extent of chromosome rearrangements responsible for the *tk* gene deletions, we analysed 9 and 32 independent *tk*-deleted clones from STK-105 and SSTK-110 cells, respectively, for heterozygosity at polymorphic microsatellite loci at variable distances from the *tk* transgene locus (Fig. 3). As many as 8 out of 9 STK-105 clones and 29 out of 32 SSTK-110 clones, which were shown to have two copies of the C57 *Dnmt1* locus, had lost heterozygosity at either all marker loci or the two more distal microsatellite loci (*D9mit297* and *D9mit256*) (Fig. 3 and Table 2, clones of classes 'A' and 'B'). This suggests that mitotic recombination had occurred between the two homologous chromosomes, with the crossovers taking place either between the centromere and *D9mit218* (class 'A', Table 2) or between *D9mit218* and the *tk* locus (class 'B', Table 2). Clones that had lost heterozygosity at all analysed markers (class 'A', Table 2) might have alternatively deleted the entire *tk*-carrying 129 chromosome 9. However, if that were so, a duplication of the C57 chromosome 9 must have occurred because these clones had two copies of the *Dnmt1* (Fig. 2b, c) and carried two chromosomes 9, as revealed by chromosome painting (not shown). A minor portion of the *tk*-deleted clones from either ES cell line (1 of 9 STK-105 and 3 of 32 SSTK-110 clones) showed only one copy (the C57 copy) of the *Dnmt1* locus (as shown in lane 2 of Fig. 2c), indicating large interstitial deletions on the 129 chromosome. This

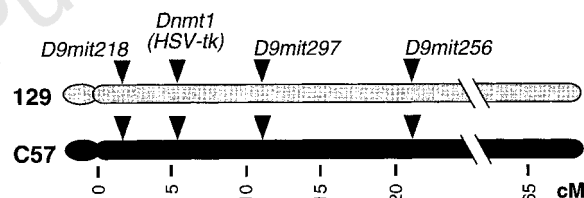


Figure 3 Diagram of mouse chromosome 9.

Table 1 Effects of DNA hypomethylation on mutation rates at *Hprt* and *HSV-tk*

ES cell line* (<i>Dnmt1</i> genotype)	Number of independent drug selections	Average number of cells screened per selection	Number of mutants per selection†	Mutation rate‡
(a) Hprt				
J1 (+/+), V18 (+/+), C/+ (+/-), C/CHIP (-/cDNA)	18	1.7×10^5	2 (1), 1 (5), 0 (12)	6.0×10^{-9}
S/S (-/-), C/C (-/-)	9	2.0×10^5	164, 73, 50, 27, 22, 20, 15, 14, 11	ND
C/C (-/-)	104	8.0×10^5	7 (1), 5 (2), 4 (1), 3 (1), 2 (3), 1 (11), 0 (85)	6.3×10^{-8}
(b) HSV-TK				
STK-105 (+/-)	9	1.4×10^5	17, 12, 11, 8, 3 (3), 2, 0	ND
	120	2.7×10^5	3 (1), 2 (1), 1 (4), 0 (114)	4.7×10^{-8}
SSTK-110 (-/-)	11	1.3×10^5	70, 46, 39, 38, 19, 15, 12, 11, 5 (2), 4	ND
	120	2.7×10^5	25 (1), 12 (1), 5 (1), 4 (1), 3 (5), 2 (5), 1 (20), 0 (86)	3.1×10^{-7}

* For the sake of simplicity, J1, V18, C/+ and C/CHIP were grouped as methylation-proficient ES cells (see the text). The number of independent drug selections performed on each individual cell line was as follows: 9 (J1), 3 (V18), 3 (C/+), 3 (C/CHIP).

† The order of numbers was arranged arbitrarily from the most to the fewest. The number in parentheses indicates the number of selection(s) giving rise to the corresponding number of drug-resistant colonies.

‡ Calculated according to the P_0 method of Luria-Delbrück fluctuation analysis¹⁰. The calculated values were corrected for plating efficiency (~40% for all the ES cell lines), but did not account for phenotypic lag in the expression of resistance to 6-TG or ganciclovir, on which differences in methylation showed little influence (not shown). Statistical significance of the differences in *Hprt* or *tk* mutation rates between methylation-proficient ES cells and their methylation-deficient counterparts was determined by the χ^2 test ($P < 0.0001$ in both cases). ND, not determined.

Table 2 Likely mechanisms for *tk* gene deletions

Independent <i>tk</i> deletion		Genotype of polymorphic loci*				Mechanism
STK-105 (<i>Dnmt1</i> ^{+/-})	SSTK-110 (<i>Dnmt1</i> ^{-/-})	<i>D9mit218</i>	<i>Dnmt1</i>	<i>D9mit297</i>	<i>D9mit256</i>	
(n = 9)	(n = 32)					
7	27	C57/C57	C57/C57	C57/C57	C57/C57	A
1	2	C57/129	C57/C57	C57/C57	C57/C57	B
1	1	C57/129	C57/-	C57/129	C57/129	C
0	1	C57/-	C57/-	C57/129	C57/129	C
0	1	C57/129	C57/-	C57/-	C57/129	C

* The genotypes were determined either by PCR (the three microsatellite loci; ref. 28), or by Southern blotting (the *Dnmt1* locus where the *tk* transgene was inserted; Fig. 2).

Mechanisms are defined as follows: A, mitotic recombination or chromosomal deletion accompanied by duplication of the remaining chromosome; B, mitotic recombination; C, interstitial deletion. See text for details.

is supported by the observation that these clones retained heterozygosity for at least one microsatellite marker that is more distal to the centromere than the *tk* locus (class 'C', Table 2). Thus, as summarized in Table 2, *tk* gene deletions, which comprised most (73%; Fig. 2d) of the observed *tk* mutations from SSTK-110 but only a minor fraction (29%; Fig. 2d) from STK-105 ES cells, were mostly due to mitotic recombination between the two homologous chromosomes or due to a chromosomal loss accompanied by duplication of the remaining chromosome. Limited deletions involving multiple loci also contributed, but to a much smaller extent.

The results presented in this paper demonstrate a mutator phenotype for *Dnmt1*-deficient ES cells. The predominance of extensive genomic changes in the observed *Hprt* and *tk* mutations in *Dnmt1*-deficient ES cells strongly suggests that DNA methylation functions in suppressing mitotic recombination and/or in contributing to faithful chromosomal segregation during mitosis. It has been reported that mitotic recombination is rarely seen in mammalian somatic cells¹¹, where genomic methylation is high¹², but relatively frequent in yeasts¹³, where no genomic methylation has been detected¹⁴. Likewise, the genomes of mammalian germ cells with obligatory meiotic recombination are hypomethylated^{15,16}. Indeed, direct evidence that methylation strongly inhibits meiotic recombination has recently been provided for *Ascomobolus immersus*¹⁷. Finally, it has been suggested that hypomethylation of the immunoglobulin genes is a prerequisite for allele-specific recombination during B-cell maturation^{18,19}. Repetitive sequences, which are highly concentrated in heterochromatin regions such as the centromere, contain most of the methylated cytosines²⁰, and have been found, in at least some instances, to be hotspots for recombination^{21,22}. It is therefore possible, that methylation of such abundant elements might represent one general mechanism for assuring genomic integrity by suppressing the potentially destabilizing effect of recombination at repetitive and transposable elements²⁰.

Our results are of potential significance for understanding the selective role of DNA hypomethylation in carcinogenesis. It has been pointed out that the spontaneous mutation rate of the normal cell is insufficient to explain carcinogenesis, suggesting that the premalignant cell must acquire a mutator phenotype²³. Because both aneuploidy and hypomethylation are observed early in the transformation process, it is tempting to speculate that genomic hypomethylation provides the incipient cancer cells with a way of acquiring more mutations by destabilizing the genome and promoting loss of heterozygosity in regions containing tumour suppressor genes. We note that hypomethylated ES cells remain largely euploid, in contrast to tumour cells. It is possible, however, that mechanisms that protect embryonic cells against genomic rearrangements are not operative in somatic cells, rendering these cells more susceptible to the destabilizing effects of hypomethylation. We also note that, depending on the particular tumour system, hypomethylation might also reduce the risk for cancer, as suggested by a previous observation that mice heterozygous for the *Dnmt1* mutation and carrying the *Min* allele were significantly protected against intestinal tumours²⁴. Subsequent evidence strongly supported the notion that in the *Min* system hypomethylation protected against tumours by an epigenetic mechanism rather than by reducing the risk of mutations²⁵. Therefore, it will be important to extend the present studies to somatic cells and assess the consequences of hypomethylation on genome stability. However, owing to the lethality of genomic demethylation to somatic cells (refs 7, 8; and our unpublished results), this will necessitate the construction of hypomorphic *Dnmt1* alleles that substantially reduce genomic methylation but do not result in cellular lethality. □

Methods

ES cell lines. J1 (ref. 7) and V18 are two independently isolated wild-type ES cell lines of inbred 129/Sv background. The C/+ (*Dnmt1*^{+/+}) ES cell line was derived from J1 and the C/C (*Dnmt1*^{-/-}) ES cell line was generated from C/+ by a

second targeting of the wild-type *Dnmt1* allele⁸. The S/S (*Dnmt1*^{S/S}) ES cell line was isolated from a *Dnmt1* homozygous mutant embryo that was derived from a targeted J1 ES clone⁸. The cDNA-rescued ES cell line (C/CHIP) was generated by inserting a *Dnmt1* cDNA into the cognate *Dnmt1* locus of C/C cells, using a protocol called CHIP⁹. ES cell line SSTK-110 (genotype: *Dnmt1*^{S/S}), which contains an integrated viral *tk* gene, was generated as follows: a 129/C57 F1 *Dnmt1*^{+/+} ES cell line that had been isolated from a cross between 129/Sv *Dnmt1*^{+/+} and C57BL/6 wild-type mice was transfected with pMT-3lox, a *Dnmt1*-targeting construct containing a CMV-hyg-*tk* cassette. A clone in which a single copy of the CMV-hyg-*tk* cassette was homologously inserted into the *Dnmt1* locus on the 129 chromosome was transfected with pMT(S)^{puro}, a *Dnmt1*-targeting vector constructed as described⁸. Of 144 puromycin-resistant clones analysed, one clone (SSTK-110) showed that the wild-type *Dnmt1* allele on the C57 chromosome was disrupted. A non-targeted clone (STK-105) from the transfection was used as a control cell line. The growth rates of the different ES cell lines were indistinguishable and their plating efficiencies were ~40%.

Spontaneous mutational analysis. Single ES cell colonies or small ES cell cultures (~50 cells) were independently expanded in the absence of feeder cells to desired cell numbers. The expanded cell populations were then trypsinized and plated at a density not greater than 10⁷ cells per 150-mm dish, in medium containing 2 µg ml⁻¹ 6-TG (Sigma; for *Hprt* selection) or 2 µM ganciclovir (for TK selection). The selection step was also performed in the absence of feeders. Counting the day of plating as day 0, media were changed on days 2, 4 and 7. On day 11, drug-resistant colonies were counted and picked for further analysis.

RT-PCR. To minimize variations in sample preparation and RT-PCR conditions, RNA samples from all the clones analysed were prepared simultaneously as described²⁶ and the same RNA preparation from each clone was used for all four RT-PCR reactions. The primers were designed according to a published mouse *Hprt* cDNA sequence²⁷. They are 5'-CCTGCTGGATTACATTAAGCACTG-3' and 5'-GTCAAGGGCATATCCAACAACAAC-3' (amplifying exons 3–8; 352-bp product); 5'-ATTAGCGATGATGAACCAAG-3' and 5'-CTGTCCATAATCAGTCCATG-3' (amplifying exon 2; 107-bp product); 5'-GGCTTCCTCCTCAGACCGCT-3' and 5'-CTCCTC GTATTTCAGATTTC-3' (amplifying exons 1–9; 738-bp product); and 5'-TTACCTCACTGCTTCCGGA-3' and 5'-TACTGGCAACATCAAC AGGA-3' (amplifying exons 1–9; 798-bp product).

Received 29 May; accepted 8 July 1998.

1. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
2. Laird, P. W. & Jaenisch, R. The role of DNA methylation in cancer genetics and epigenetics. *Annu. Rev. Genet.* **30**, 441–464 (1996).
3. Lengauer, C., Kinzler, K. W. & Vogelstein, B. DNA methylation and genetic instability in colorectal cancer cells. *Proc. Natl. Acad. Sci. USA* **94**, 2545–2550 (1997).
4. Jeanpierre, M. et al. An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome. *Hum. Mol. Genet.* **2**, 731–735 (1993).
5. Ji, W. et al. DNA demethylation and pericentromeric rearrangements of chromosome 1. *Mutat. Res.* **379**, 33–41 (1997).
6. Haaf, T. The effects of 5-azacytidine and 5-azadeoxycytidine on chromosome structure and function: implications for methylation-associated cellular processes. *Pharmacol. Ther.* **65**, 19–46 (1995).
7. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
8. Lei, H. et al. *De novo* DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195–3205 (1996).
9. Tucker, K. L. et al. Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes Dev.* **10**, 1008–1020 (1996).
10. Kendal, W. S. & Frost, P. Pitfalls and practice of Luria–Delbrück fluctuation analysis: a review. *Cancer Res.* **48**, 1060–1065 (1988).
11. Shulman, M. J., Collins, C., Connor, A., Read, L. R. & Baker, M. D. Interchromosomal recombination is suppressed in mammalian somatic cells. *EMBO J.* **14**, 4102–4107 (1995).
12. Jaenisch, R. DNA methylation and imprinting: why both? *Trends Genet.* **13**, 323–329 (1997).
13. Petes, T. D., Malone, R. E. & Symington, L. S. In *The Molecular and Cellular Biology of the Yeast Saccharomyces* (eds Broach, J. R., Pringle, J. R. & Jones, E. W.) 407–521 (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1991).
14. Proffitt, J. H., Davie, J. R., Swinton, D. & Hattman, S. 5-Methylcytosine is not detectable in *Saccharomyces cerevisiae* DNA. *Mol. Cell. Biol.* **4**, 985–988 (1984).
15. Monk, M., Boubelik, M. & Lehnert, S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* **99**, 371–382 (1987).
16. Sanford, J. P., Clark, H. J., Chapman, V. M. & Rossant, J. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes Dev.* **1**, 1039–1046 (1987).
17. Maloel, L. & Rossignol, J.-L. Suppression of crossing-over by DNA methylation in *Ascomobolus*. *Genes Dev.* **12**, 1381–1389 (1998).
18. Hsieh, C.-L. & Lieber, M. R. CpG methylated minichromosomes become inaccessible for V(D)J recombination after undergoing replication. *EMBO J.* **11**, 315–325 (1992).
19. Engler, P., Weng, A. & Storb, U. Influence of CpG methylation and target splicing on V(D)J recombination in a transgenic substrate. *Mol. Cell. Biol.* **13**, 571–577 (1993).

20. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
21. Wahls, W. P., Wallace, L. J. & Moore, P. D. Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell* **60**, 95–103 (1990).
22. Shiroishi, T., Koide, T., Yoshino, M., Sagai, T. & Moriawaki, K. Hotspots of homologous recombination in mouse meiosis. *Adv. Biophys.* **31**, 119–132 (1995).
23. Loeb, L. A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075–3079 (1991).
24. Laird, P. W. *et al.* Suppression of intestinal neoplasia by DNA hypomethylation. *Cell* **81**, 197–205 (1995).
25. Jackson-Grusby, L., Laird, P. W., Magge, S. N., Moeller, B. J. & Jaenisch, R. Mutagenicity of 5-aza-2'-deoxycytidine is mediated by the mammalian DNA methyltransferase. *Proc. Natl. Acad. Sci. USA* **94**, 4681–4685 (1997).
26. Beard, C., Li, E. & Jaenisch, R. Loss of methylation activates *Xist* in somatic but not in embryonic cells. *Genes Dev.* **9**, 2325–2334 (1995).
27. Melton, D. W., Konecki, D. S., Brennand, J. & Caskey, C. T. Structure, expression, and mutation of the hypoxanthine phosphoribosyltransferase gene. *Proc. Natl. Acad. Sci. USA* **81**, 2147–2151 (1984).
28. Dietrich, W. F. *et al.* A genetic map of the mouse with 4,006 simple sequence length polymorphisms. *Nature Genet.* **7**, 220–245 (1994).

Acknowledgements. We thank P. P. Lee for pMT-3lox; W. M. Rideout III, and J. T. Lee for critical reading of the manuscript; and W. Thilly for helpful discussion. This work was supported by a grant from the Swedish Medical Research Council (U.P.), postdoctoral fellowships from the Ann Fuller Fund (R.Z.C.) and the Cancer Research Fund of the Damon Runyon–Walter Winchell Foundation (L.J.-G.), and by the National Institutes of Health (R.Z.C., L.J.-G. and R.J.).

correspondence and requests for materials should be addressed to R. J. (e-mail: jaenisch@wi.mit.edu).

RNA polymerase II is an essential mRNA polyadenylation factor

Yutaka Hirose* & James L. Manley

Department of Biological Sciences, Columbia University, New York, New York 10027, USA

Production of messenger RNA in eukaryotic cells is a complex, multistep process. mRNA polyadenylation, or 3' processing, requires several protein factors, including cleavage/polyadenylation-specificity factor (CPSF), cleavage-stimulation factor, two cleavage factors and poly(A) polymerase (reviewed in refs 1, 2). These proteins seem to be unnecessary for other steps in mRNA synthesis such as transcription and splicing, and factors required for these processes were not considered to be essential for polyadenylation. Nonetheless, these reactions may be linked so that they are effectively coordinated *in vivo*^{3–9}. For example, the CTD carboxy-terminal domain of the largest subunit of RNA polymerase II (RNAP II) is required for efficient splicing and polyadenylation *in vivo*⁸, and CPSF is brought to a promoter by the transcription factor TFIID and transferred to RNAP II at the time of transcription initiation⁹. These findings suggest that polyadenylation factors can be recruited to an RNA 3'-processing signal by RNAP II, where they dissociate from the polymerase and initiate polyadenylation. Here we present results that extend this model by showing that RNAP II is actually required, in the absence of transcription, for 3' processing *in vitro*.

We previously investigated a possible energy requirement for the first step of the polyadenylation reaction (the cleavage step), and found that creatine phosphate could function as a necessary cofactor but that hydrolysis of the high-energy phosphate bond did not occur¹⁰. Part of the evidence for this was that arginine phosphate could substitute for creatine phosphate in the reaction, even though it is not a substrate for creatine phosphokinase. One explanation for this could be that phosphoguanidino compounds can function as molecular mimics, substituting for an unknown phosphoprotein. As phosphoarginine is rarely found in proteins¹¹, we tested whether cleavage could be activated by other phospho-amino acids in a reconstituted processing reaction without creatine

phosphate and containing an SV40 late pre-mRNA substrate, which can be cleaved efficiently by a combination of CPSE, cleavage-stimulation factor (CstF), and the cleavage factors CFI and CFII (ref. 12). Figure 1a indicates that increasing concentrations of phosphoserine (lanes 7–11) resulted in cleavage that was as efficient as with creatine phosphate (lanes 2–6). AMP was unable to activate cleavage (Fig. 1a, lanes 12–16), as were many other phosphate compounds¹⁰.

We next tested whether certain specific phosphoproteins could participate in the 3'-cleavage reaction in the absence of creatine phosphate. Poly(A) polymerase (PAP) is required for cleavage of most pre-mRNAs^{13,14} and can be extensively phosphorylated in its carboxy-terminal regulatory region^{15,16}, but was unable to activate cleavage in our assay in the absence of creatine phosphate (results not shown). The SR family of splicing factors¹⁷ are also heavily phosphorylated in their C-terminal domains, and although their primary function is in splicing they may in at least one instance play a role in polyadenylation¹⁸; however, we found that purified SR proteins were unable to reconstitute cleavage (results not shown,

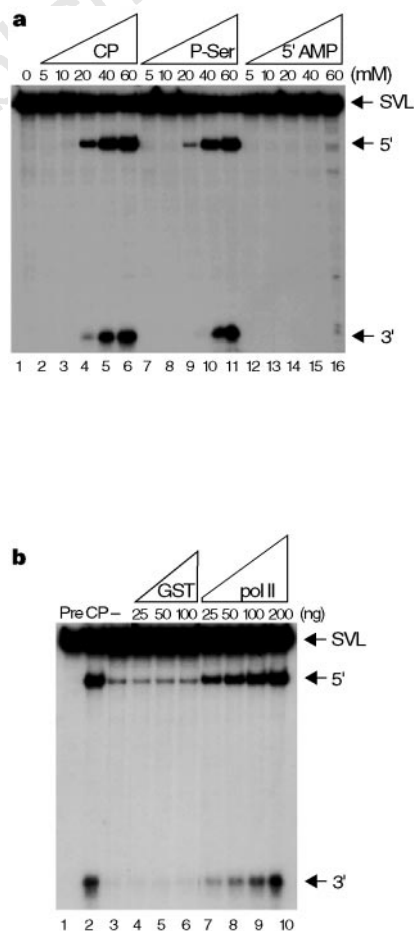


Figure 1 Phosphoserine and RNA polymerase II activate pre-mRNA 3' cleavage. **a**, *In vitro* cleavage reactions were carried out with ³²P-labelled SV40 late pre-mRNA substrate in a standard cleavage reaction mixture containing CPSF, CstF, CFI, CFII and 2 mM EDTA for 90 min at 30 °C without phosphate compound (lane 1), or with increasing amounts (5–60 mM) of creatine phosphate (CP) (lanes 2–6), phosphoserine (P-Ser) (lanes 7–11), or 5' AMP (lanes 12–16). RNA products were isolated and fractionated on 5% polyacrylamide, 8.3 M urea gels. **b**, 3' cleavage of SV40 pre-mRNA was assayed as in **a**, except that EDTA was replaced by 0.5 mM MgCl₂ in the presence of 40 mM CP (lane 2), no addition (lane 3), increasing amounts (25–100 ng) of purified GST (lanes 4–6), or increasing amounts (25–200 ng) of highly purified HeLa RNAP II (lanes 7–10). Unprocessed substrate RNA (Pre) is shown in lane 1. The positions of precursor RNA (SVL), 5' and 3' cleavage products (denoted as 5' and 3') are indicated.

* Present address: Department of Biophysics, Cancer Research Institute, Kanazawa University, 13-1, Takaramachi, Kanazawa, Ishikawa 920, Japan.

and Fig. 2b). We next tested RNAP II. The CTD of RNAP II associates with 3'-processing factors, apparently recruiting them to nascent mRNA; the CTD can also be heavily phosphorylated *in vivo*¹⁹. Remarkably, we found that RNAP II (Fig. 1b, pol II; lanes 7–10) was able, in a concentration-dependent way, to reconstitute processing to the level observed in the presence of creatine phosphate (Fig. 1b, lane 2). These results are consistent with the hypothesis that creatine phosphate acts in polyadenylation by mimicking RNAP II.

We confirmed that RNAP II is a necessary polyadenylation factor in the absence of creatine phosphate in an additional reconstitution experiment which differs from the above assay in several ways. First, we used another pre-mRNA, adenovirus L3 (L3) pre-mRNA. This allowed us to extend our results to a second RNA and to show that our findings were not related to the unusual PAP-independence of SV40 late pre-mRNA. Second, because the RNAP II sample used above was not homogeneous despite being highly purified, we purified RNAP II to apparent homogeneity according to ref. 20. This procedure allowed us to separate the IIO form, which contains a hyperphosphorylated CTD, from the IIA isoform, which has a hypophosphorylated CTD: aliquots of the two preparations were resolved by SDS-PAGE and the gel was silver-stained (Fig. 2a). These samples were then used in 3'-processing reactions containing L3 RNA and CPSE, CstF, CFI, CFII and PAP (Fig. 2b). Strikingly, both RNAP IIA (lanes 3–5) and IIO (lanes 6–8) reconstituted efficient cleavage, indicating not only that RNAP II is an essential polyadenylation factor, but also that CTD phosphorylation, as suggested by our results with creatine phosphate, may be unnecessary (but see later). The concentrations of RNAP II needed for

efficient cleavage were low (5 ng, or ~1 nM), comparable to those of the characterized polyadenylation factors required for optimal processing, and are consistent with RNAP II playing an integral, stoichiometric role in the cleavage reaction. RNAP II does not appear to be involved in the second, poly(A)-synthesis phase of the reaction (data not shown).

Although the behaviour of RNAP IIA and IIO was the same in these assays, we nonetheless investigated whether the CTD was itself sufficient to reconstitute 3' cleavage, by expressing a glutathione S-transferase(GST)-CTD fusion protein in *Escherichia coli*, purifying it to near-homogeneity, then phosphorylating and rephosphorylating it (see Methods). The recombinant proteins were analysed by SDS-PAGE (Fig. 3a) and equivalent concentrations tested for their ability to reconstitute processing of L3 pre-mRNA (Fig. 3b). We found that both the unphosphorylated (Fig. 3b, lanes 3–6) and phosphorylated (Fig. 3b, lanes 7–10) forms of the fusion protein activated cleavage

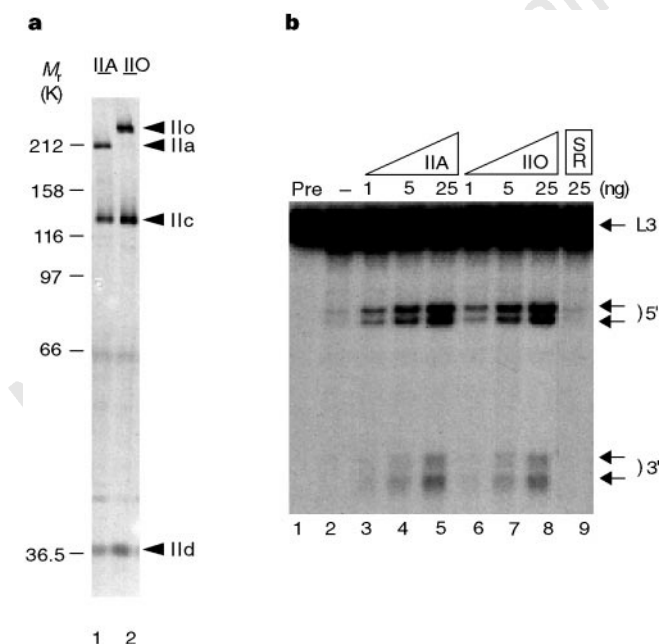


Figure 2 Both IIA and IIO isoforms of RNA polymerase II can activate 3' cleavage. **a**, Silver staining of 7.5% SDS-PAGE gel containing homogeneously purified hypo- (lane 1) and hyper- (lane 2) phosphorylated RNAP II. Positions of size-marker proteins are indicated on the left. Migration of hypo- (IIA) or hyper- (IIO) phosphorylated largest subunit, second- (IIC) and third- (IID) largest subunits of RNAP II are indicated on the right. **b**, 3' cleavage of adenovirus L3 pre-mRNA substrate was assayed under standard conditions as for Fig. 1, except that EDTA was replaced by 2 mM MgCl₂ and reaction mixtures also contained 1 mM 3'-dATP and 10 ng recombinant bovine PAP II. Reactions were done without additional protein (lane 2), or with increasing amounts (1–25 ng) of hypo- (lanes 3–5) or hyper- (lanes 6–8) phosphorylated RNAP II, or with 25 ng of HeLa purified SR proteins (lane 9). The position of precursor RNA (L3), and 5' and 3' cleavage products (denoted as 5' and 3') are indicated.

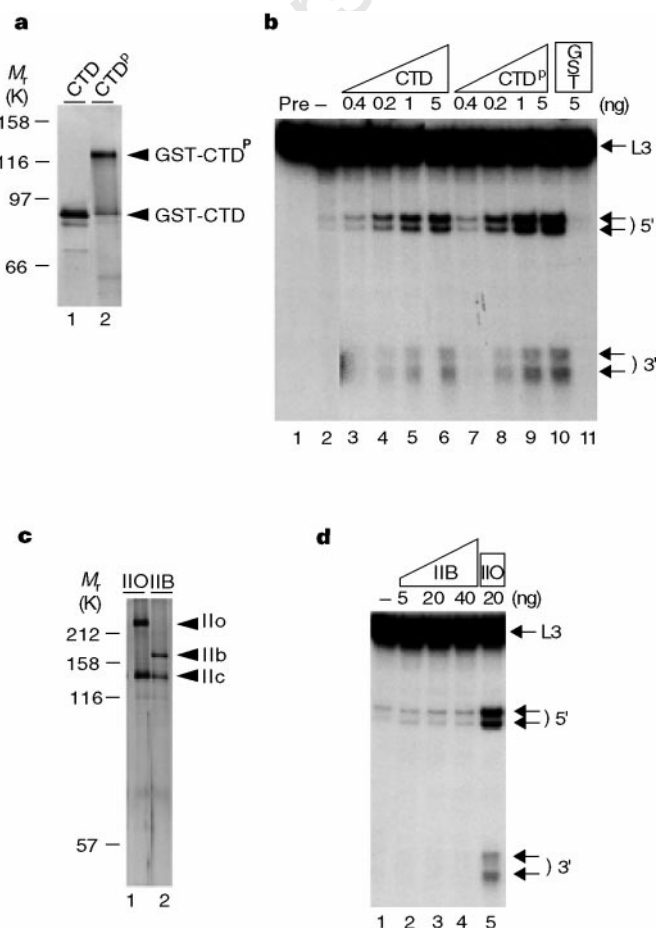


Figure 3 The CTD is necessary and sufficient to reconstitute 3' cleavage. **a**, Silver staining of 7.5% SDS-PAGE gel containing non-phosphorylated (lane 1) or *in vitro* phosphorylated (lane 2) GST-CTD purified from *E. coli*. Positions of size-marker proteins are indicated on the left. Migration of the non-phosphorylated and phosphorylated GST-CTD proteins are indicated. **b**, 3' cleavage of L3 substrate was assayed under the same conditions as for Fig. 2b, without additional protein (lane 2), or with increasing amounts (0.04–5 ng) of non-phosphorylated (lanes 3–6) or phosphorylated (lanes 7–10) GST-CTD, or with 5 ng of purified GST alone (lane 11). Positions of the RNA products are indicated. **c**, Silver staining of 10% SDS-PAGE gel containing purified RNAP IIO (lane 1) and RNAP IIB (lane 2). Positions of size-marker proteins are indicated on the left. Migration of the hyperphosphorylated largest subunit (IIO), the CTD-less largest subunit (IIB), and the second-largest subunit (IIC) of RNAP II are indicated on the right. **d**, 3' cleavage of the L3 substrate was assayed, under the same conditions as in Fig. 2b, without additional protein (lane 1), or with increasing amounts (5–40 ng) of RNAP IIB (lanes 2–4) or with 20 ng of RNAP IIO (lane 5).

in a concentration-dependent way, whereas GST alone (Fig. 3b, lane 11) was inactive. The amount of GST-CTD required was small, with cleavage being detectable with as little as 40 pg. The concentrations of RNAP II (Fig. 2b) and GST-CTD required for optimal cleavage (1–5 nM) were almost identical, strongly suggesting that the CTD is indeed the active component. By contrast with the IIO and IIA forms of RNAP II, which behaved identically, phosphorylated GST-CTD was significantly more active (~5-fold) than the unphosphorylated form.

We next tested purified RNAP IIB, a proteolytic product lacking the CTD (gift from X. Sun and D. Reinberg), to determine whether the CTD was also necessary for RNAP II-activated cleavage. A silver-

stained SDS-polyacrylamide gel of the largest subunits is shown in Fig. 3c; Fig. 3d shows that RNAP IIB was unable to activate cleavage. Taken together, our results indicate that the CTD is both necessary and sufficient for activation of 3' processing by RNAP II.

Although we have shown that RNAP II can very effectively reconstitute 3' cleavage when mixed with the appropriate factors, it remained possible that RNAP II might be filling a role naturally played by some other factor missing from our reconstituted processing reactions. To exclude this possibility, we investigated whether RNAP II was required for cleavage in an unfractionated nuclear extract by testing the effect of immunodepletion of RNAP II on 3' processing and then of adding purified RNAP II back to the assay. For immunodepletion, we used a monoclonal antibody, 7G6, directed against the CTD²¹, and monitored the efficiency of RNAP II depletion by western blotting. As shown in Fig. 4a, we were able to reduce the amount of RNAP II in the depleted extract (lane 2) to ~20% of that in mock-depleted (lane 1) or untreated (lane 3) extracts, but no further; the IIA isoform was more easily removed than IIO. Western blots also confirmed that there was no depletion of polyadenylation factors (Fig. 4a). Increasing concentrations of the mock- and RNAP II-depleted extracts were then used in 3'-cleavage reactions with L3 pre-mRNA (Fig. 4b): the activity of RNAP II-depleted extract (lanes 5–7) was significantly reduced relative to mock-depleted extract (lanes 2–4) and the degree of inhibition (4–5-fold) correlated well with the extent of RNAP II depletion.

To verify that the inhibition of cleavage was due to removal of RNAP II and not of an associated protein, we added highly purified RNAP IIA, IIO or phosphorylated GST-CTD back to the immunodepleted extract (Fig. 4c). None of the proteins had a large effect on the activity of the mock-depleted extract, although the highest concentration of RNAP IIO (lane 6), and to a lesser extent IIA (lane 4), did increase cleavage slightly (by a factor of less than 2). This small enhancement was not seen with any of the other polyadenylation factors (results not shown), suggesting that RNAP II may be the limiting factor for 3' cleavage in nuclear extracts. All three proteins restored significant cleavage in the RNAP II-depleted extract (Fig. 4c: compare lane 8 with lanes 9–13); RNAP IIO had the most striking effect, with 25 ng restoring processing in the RNAP II-depleted extract to essentially the same as that in RNAP IIO-supplemented, mock-depleted extract (Fig. 4c: compare lanes 6 and 12). RNAP IIA (Fig. 4c, lanes 9, 10) and phosphorylated CTD (Fig. 4c, lane 13) were less effective than RNAP IIO, in contrast to their behaviour in assays employing purified factors (Figs 2, 3): this may reflect the ability of these proteins to interact differentially with other factors present in the nuclear extract but not in the reconstituted reactions, which might in turn influence their ability to participate in 3' cleavage. In agreement with the reconstitution experiments (Fig. 3d), RNAP IIB was unable to restore cleavage in RNAP II-depleted extract (Fig. 4d).

Our results have shown that RNAP II, in the absence of transcription, is a necessary factor for pre-mRNA 3' processing *in vitro*. We propose that the CTD of RNAP II participates in the formation of a stable, catalytically active processing complex, extending previous models in which RNAP II recruits processing factors to the polyadenylation signal during transcription, thereby helping to define the poly(A) site on pre-mRNA^{8,9}. In these earlier models, RNAP II then dissociates from the processing factors and continues transcription, perhaps now in termination mode, and did not play any direct part in the cleavage reaction; our data, however, indicate that RNAP II participates directly and actively in processing, probably by interaction of its CTD with one or more processing factors—both CPSF and CstF can associate with the CTD, perhaps reflecting a direct interaction with one factor, CstF-50 (ref. 8).

We also found that both the hypo- and hyperphosphorylated forms of RNAP II could reconstitute cleavage. We initially suspected that RNAP IIO might be the active form, given the creatine

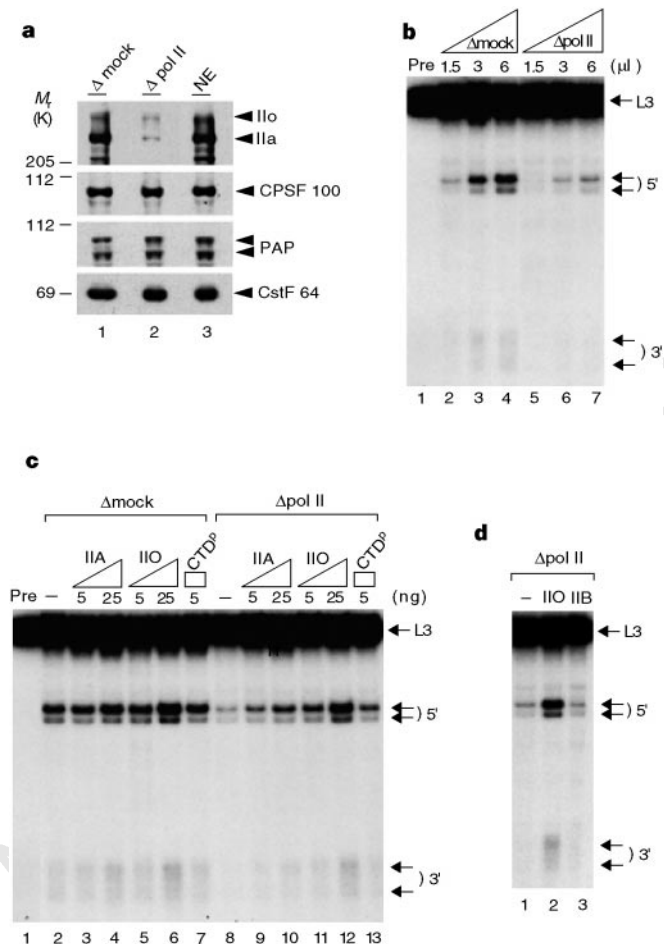


Figure 4 RNA polymerase II is required for 3' cleavage in nuclear extract. **a**, Western blot of HeLa cell nuclear extract (NE) depleted with anti-CTD antibody (lane 2), mock-depleted (lane 1) or non-treated (lane 3). Equivalent amounts of each extract were fractionated by SDS-PAGE and probed with anti-CTD antibody (top panel) or with antibodies directed against the indicated polyadenylation factors (bottom panels). The positions of size markers are shown on the left, and of the hypo- (IIa) and hyper- (IIO) phosphorylated forms of the RNAP II largest subunit and of the indicated polyadenylation factors on the right. **b**, 3' cleavage of L3 pre-mRNA substrate was carried out with increasing amounts (1.5–6 μ l) of either mock-depleted (lanes 2–4) or RNAP II-depleted (lanes 5–7) nuclear extract. **c**, 3' cleavage of L3 pre-mRNA substrate was done with 5 μ l mock-depleted (lanes 2–7) or RNAP II-depleted extract (lanes 8–13) in the presence of the indicated amounts of purified RNAP IIA (lanes 3, 4, 9 and 10), RNAP IIO (lanes 5, 6, 11 and 12), and phosphorylated GST-CTD (lanes 7 and 13). Cleavage in the RNAP II-depleted extract was reduced by a factor of 4 relative to mock-depleted extract. **d**, Cleavage of the L3 substrate was done with 4 μ l RNAP II-depleted extract in the absence of additional protein (lane 1), or with 20 ng RNAP IIO (lane 2) or 20 ng RNAP IIB (lane 3). Positions of RNA products are indicated.

phosphate results and that IIO is the elongating form of the polymerase¹⁹, and so probably the one to encounter the poly(A) signal in nascent RNA. However, the IIA form might also be functional, which is consistent with the indistinguishable interaction of CstF and CPSF with phosphorylated and unphosphorylated GST-CTD⁸. But as RNAP IIO was more active in depleted nuclear extract, this may be the physiological 'RNA-processing' form of the enzyme.

Our results have revealed a link between transcription and polyadenylation that is much closer than was previously thought. They open up new possibilities for gene control and for understanding the timing and mechanism of 3'-processing events. For example, the direct involvement of RNAP II in 3' cleavage may indicate that the reaction must occur immediately after the poly(A) site is transcribed. Alternatively, RNAP II may play an essential but transient role, perhaps aiding formation of a conformation necessary for processing. Lastly, transcription could continue while the CTD remains in contact with processing factors, resulting in a looping-out of the newly synthesized RNA. Phosphorylation of the CTD is not essential, but rather it enhances the efficiency of cleavage, suggesting that post-translational modification of this domain could regulate 3'-end formation. Also, specific proteins may influence the interaction of the CTD with polyadenylation factors. In any case, our findings point to an unexpected integration of nuclear processes and raise new questions about the mechanism and regulation of gene expression. □

Methods

Proteins. Recombinant bovine PAP II was purified from *E. coli* or baculovirus-infected SF9 cells¹⁵. All other 3' processing factors were purified from HeLa cell nuclear extract²², using the initial steps of purification described^{12,23}. CPSF was further purified by chromatography on a TSK-heparin-5PW column; CFI and CFII were further purified from Mono-S fractions by re-chromatography on Mono-S columns. These preparations were all free of detectable RNAP II, except for one (CPSF), which contained trace (sub-nanogram) amounts, thus possibly explaining the low levels of cleavage observed in some experiments in the absence of added RNAP II. RNAP II (in Fig. 1) was purified from HeLa cell nuclear-extract pellets²⁴. Purification and separation of the two RNAP II isoforms (IIA and IIO) from nuclear-extract pellets was done as described²⁰. SR proteins were purified from nuclear extract²⁵. The murine GST-CTD fusion protein was expressed in *E. coli* and purified as described²⁶, except that the protein was further purified on a Superose-12 column.

In vitro phosphorylation of GST-CTD. 10 µg purified GST-CTD was incubated with 3 mg of nuclear extract in 20 mM HEPES-NaOH, pH 7.9, 8 mM MgCl₂, 1 mM ATP, 20 mM CP, 5 mM β-glycerol phosphate, 1 mM DTT and protease inhibitors, for 5 h at 30°C. Reactions were stopped by adding 10 mM EDTA and 60 µl glutathione-agarose beads (Sigma) equilibrated with buffer D²² containing 50 mM (NH₄)₂SO₄. The mixture was rocked for 1 h at 4°C and beads were pelleted by brief centrifugation and washed 5 times with 1 ml buffer A (buffer D containing 50 mM Tris-HCl, pH 7.9, instead of HEPES) containing 1 M NaCl and twice with 1 ml buffer A containing 1% Triton X-100 and 50 mM (NH₄)₂SO₄. Bound GST-CTD was eluted with buffer A plus 15 mM glutathione, and dialysed into buffer D containing 50 mM (NH₄)₂SO₄. The degree of phosphate incorporation into GST-CTD was estimated to be ~15 mol phosphate per mol GST-CTD.

3'-cleavage assays. ³²P-labelled pre-mRNA substrates were prepared as described¹⁰. A 12.5-µl reaction mixture with purified factors for the SV40 substrate contained 15 ng CstF, 150 ng CPSF, 300 ng CFI, 80 ng CFII, 500 ng BSA, 0.2–0.5 ng labelled RNA, 250 ng tRNA, 0.25 U RNasin (Promega), 9.6 mM HEPES-NaOH (pH 7.9), 9.6% glycerol, 24 mM (NH₄)₂SO₄, 0.12 mM EDTA, 0.24 mM DTT, 2.5% (w/v) polyvinyl alcohol, 0.24 mM PMSF, and the indicated amount of EDTA or MgCl₂. Cleavage assays for the L3 substrate with purified factors were done in the same way, except that reaction mixtures also contained 1 mM 3' dATP, 2 mM MgCl₂, and 10 ng recombinant PAP II. Cleavage assays in HeLa nuclear extract were carried out as above, except that nuclear extract replaced the purified factors and MgCl₂ was omitted. All reaction mixtures were incubated for 60–90 min at 30°C, and RNA products

were fractionated on 5% polyacrylamide, 8.3M urea gels. Cleavage efficiency was quantified by PhosphorImager (Molecular Dynamics).

Immunodepletion. HeLa cell nuclear extract for immunodepletion was prepared by extracting isolated nuclei in ~240 mM (NH₄)₂SO₄ as described²⁷ and adjusting to 500 mM (NH₄)₂SO₄. Extracts were clarified by centrifugation at 30,000g for 20 min and then passed through a 0.8-µm filter. 120 µg of monoclonal anti-CTD antibody (7G; ref. 21) was crosslinked to 90 µl protein G-Sepharose beads (Sigma)²⁸. Crosslinked beads were equilibrated to buffer A containing 500 mM (NH₄)₂SO₄ and divided into three aliquots. Three tubes of beads in the same buffer were prepared for mock depletion; 300 µl of nuclear extract was added to the first batch of beads and rocked for 2 h at 4°C. After a brief spin, the supernatant was transferred to the next tube and the process repeated. After three depletions, supernatants were dialysed into buffer D containing 50 mM (NH₄)₂SO₄. Immunodepletion experiments were repeated 2–4 times with 2 separate depleted extracts, with variations of no more than 20%.

Received 17 March; accepted 29 June 1998.

- Wahle, E. & Keller, W. The biochemistry of polyadenylation. *Trends Biochem. Sci.* **21**, 247–250 (1996).
- Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
- Niwa, M., Rose, S. D. & Berget, S. M. *In vitro* polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev.* **4**, 1552–1559 (1990).
- Gunderson, S. I. *et al.* The human U1A snRNP protein regulates polyadenylation via direct interaction with poly(A) polymerase. *Cell* **76**, 531–541 (1994).
- Lutz, C. S. *et al.* Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency *in vitro*. *Genes Dev.* **10**, 325–337 (1996).
- Yuryev, A. *et al.* The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc. Natl Acad. USA* **93**, 6975–6980 (1996).
- Mortillaro, M. J. *et al.* A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. *Proc. Natl Acad. USA* **93**, 8253–8257 (1996).
- McCracken, S. *et al.* The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357–361 (1997).
- Dantonel, J. C., Murthy, K. G. K., Manley, J. L. & Toral, L. CPSF links transcription and mRNA 3' end formation. *Nature* **389**, 399–402 (1997).
- Hirose, Y. & Manley, J. L. Creatine phosphate, not ATP, is required for 3' end cleavage of mammalian pre-mRNA *in vitro*. *J. Biol. Chem.* **272**, 29636–29642 (1997).
- Matthews, H. R. Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade. *Pharmacol. Ther.* **67**, 323–350 (1995).
- Takagaki, Y., Ryner, L. C. & Manley, J. L. Four factors are required for 3' end cleavage of pre-mRNAs. *Genes Dev.* **3**, 1711–1724 (1989).
- Takagaki, Y., Ryner, L. C. & Manley, J. L. Separation and characterization of a poly(A) polymerase and a cleavage specificity factor required for pre-mRNA polyadenylation. *Cell* **52**, 731–742 (1988).
- Christofori, G. & Keller, W. 3' cleavage and polyadenylation of mRNA precursors *in vitro* requires a poly(A) polymerase, a cleavage factor, and a snRNP. *Cell* **54**, 875–889 (1988).
- Colgan, D. F., Murthy, K. G. K., Prives, C. & Manley, J. L. Cell-cycle related regulation of poly(A) polymerase by phosphorylation. *Nature* **384**, 282–285 (1996).
- Colgan, D. F., Murthy, K. G. K., Zhao, W., Prives, C. & Manley, J. L. Inhibition of poly(A) polymerase requires p34cdc2/cyclin B phosphorylation of multiple consensus and non-consensus sites. *EMBO J.* **17**, 1053–1062 (1998).
- Manley, J. L. & Tacke, R. SR proteins and splicing control. *Genes Dev.* **10**, 1569–1579 (1996).
- Lou, H., Gagel, R. F. & Berget, S. M. An intron enhancer recognized by splicing factors activates polyadenylation. *Genes Dev.* **10**, 208–219 (1996).
- Dahmus, M. E. Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *J. Biol. Chem.* **271**, 19009–19012 (1996).
- Lu, H., Flores, O., Weinmann, R. & Reinberg, D. The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex. *Proc. Natl Acad. USA* **88**, 10004–10008 (1991).
- Besse, S., Vigneron, M., Pichard, E. & Puvion-Dutilleul, F. Synthesis and maturation of viral transcripts in herpes simplex virus type 1 infected HeLa cells: the role of interchromatin granules. *Gene Expr.* **4**, 143–161 (1995).
- Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11**, 1475–1489 (1983).
- Takagaki, Y., Manley, J. L., MacDonald, C. C., Wilusz, J. & Shenk, T. A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev.* **4**, 2112–2120 (1990).
- Reinberg, D. & Roeder, R. G. Factors involved in specific transcription by mammalian RNA polymerase II. *J. Biol. Chem.* **262**, 3310–3321 (1987).
- Zahler, A. M., Lane, W. S., Stolk, J. A. & Roth, M. B. SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev.* **6**, 837–847 (1992).
- Peterson, S. R., Dvir, A., Anderson, C. W. & Dynan, W. S. DNA binding provides a signal for phosphorylation of the RNA polymerase II heptapeptide repeats. *Genes Dev.* **6**, 426–438 (1992).
- Flaherty, S. M., Fortes, P., Izaurralde, E., Mattaj, J. W. & Gilmartin, G. M. Participation of the nuclear cap binding complex in pre-mRNA 3' processing. *Proc. Natl Acad. USA* **94**, 11893–11898 (1997).
- Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1988).

Acknowledgements. We thank C. Li for the RNAP II used in Fig. 1 and for discussion; X. Sun and D. Reinberg for purified RNAP IIB; K. G. K. Murthy for recombinant PAP and antibodies; R. Tacke for purified SR proteins; W. S. Dynan for the GST-CTD plasmid; L. Toral, C. Kedinger and M. Vigneron for mAb 7G5; X. H. Shi, Y. Chen and Y. Sun for help in preparing nuclear extracts; and J. Dahlberg for suggesting that arginine phosphate might function by mimicking a phosphoarginine-containing protein. This work was supported by a grant from the NIH to J.M.; Y.H. was partly supported by the Japanese Ministry of Education, Science, and Culture.

Correspondence and requests for materials should be addressed to J.L.M. (e-mail: jlm2@columbia.edu)

Equipment for field and lab

Accessories and refinements presented here include a blood pressure computer for small animals, a spectrophotometric pipette, pH measuring instruments, topographic map software and a clinical chemistry system.

Protector

From Labconco

A controlled atmosphere glove box that provides a leak-free protective barrier

Designed to deal with environment-sensitive materials, this glove box uses a welded, type 304 stainless steel liner or a one-piece moulded fibreglass liner. Protector provides a protective barrier for applications involving organometallic, oxygen-sensitive or moisture-sensitive materials (it is factory tested for leaks greater than 1×10^{-6} ml s⁻¹ or 31.55 ml yr⁻¹). Four manually controlled valves regulate gas input and exhaust. The large 0.95-cm thick, laminated safety glass window is 91-cm wide by 71-cm high, and angled at 10° to minimize reflections. The fluorescent lamp mounted exterior to the cabinet is said to provide glare-free interior lighting. Each box has two three-wire, 115-V receptacles mounted inside, as well as one auxiliary receptacle on the outside. These are 20-cm diameter glove ports spaced at 43-cm centres. Double grooved rings allow glove changes without disturbing chamber integrity. One pair of 0.015-gauge neoprene gloves, size 9-3/4 is included.

Reader Enquiry No. 100

Traceable clock

From Markson

A traceable clock that receives a signal directly from the NIST atomic clock

Accurate to better than one millionth of a second per year, the US National Institute of Standards and Technology radio signal automatically sets and precisely displays the time-of-day (to the second), a.m./p.m., month, date and day of the week. It also adjusts automatically for daylight savings time, leap year and Earth rotation corrections. The Traceable Clock features electroluminescent backlighting for comfortable night viewing, a settable alarm with adjustable volume and a 1.2-cm LCD display mounted in a compact 2 × 10 × 3.8-cm case. This system should prove useful for standardizing clocks and computers, as



Clock watching — this traceable clock keeps time.

well as supporting accurate laboratory documentation.

Reader Enquiry No. 101

SpectroPipetter

From Ocean Optics

Direct spectrophotometric measurements within the pipette sampling device

The plunger, which is the actual optical fibre, is connected to a spectrometer carrying light to and from the sample, and retrieves the sample, as well. The tip of the pipette includes a proprietary micro-sample cell that acts as an optical waveguide for aqueous sample solutions. The instrument/liquid handler couples to the company's high-sensitivity fibre-optic spectrometers and compact light sources. It features rapid measurement directly in the pipette. The flexible, 10-mm path length is useful for applications requiring very small sample volumes (only 2 µl), such as protein chemistry, molecular biology, forensics and toxicology.

Reader Enquiry No. 102

Alcyon 300 and 300i

From Abbott Laboratories

Benchtop clinical chemistry systems

These Alcyon analysers are intended for labs that need small, fast systems for routine and specialized chemistry testing (especially those running fewer than 1,000 tests per day or 100,000 annually). Both feature a 49-assay menu of liquid, ready-to-use reagents, an automated cuvette handler, 39 sample holders, 24 reagent positions and the ability to process up to 300 photometric tests per hour. The 300i model adds an integrated ISE (chloride, potassium, sodium) module to further reduce labour and instrument costs. These new models are said to provide faster testing capabilities and an improved graphical user interface with a colour touchscreen. The systems can use Abbott and non-Abbott reagents. There is also a cuvette management and processing centre that features an on-board capacity of 223 cuvettes, providing maximum walk-away time.

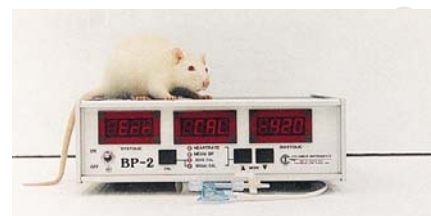
Reader Enquiry No. 103

BP-2

From Columbus Instruments

A physiological blood pressure computer for small and large animals

This instrument utilizes a cannulation method and a miniature blood pressure



Columbus' blood pressure computer: BP-2.

transducer. It measures and displays systolic, diastolic and mean blood pressures, as well as heart rate in the range of 0–600 b.p.m., which is needed for monitoring small rodents. It is equipped with three bright 1.9-cm LED displays, which are visible even across a brightly lit room. An RS-232 communication option allows the transfer of calculated blood pressure values to the computer. The analog voltage, which represents the instantaneous blood pressure, is also accessible on the back panel for an external chart recorder.

Reader Enquiry No. 104

Field equipped

TopoScout

From Maptech

GPS-compatible, topographic maps from the United States Geological Survey (USGS)

While this system is aimed at recreational users, it may prove useful to field scientists, as well. The program should help researchers plan field excursions at a computer in the office, en route or in the field. Users can print their own version of a map, record trails, mark and measure distance and area with just a click of a mouse. Maps can be viewed at four zoom levels, with path elevation and line of sight profiling, together with an extensive place finder. Users can easily enter GPS coordinates in latitude and longitude, UTM or MGRS and bookmark the map for instant retrieval. System requirements include Windows95, 3.1 or NT; a 486 CPU or better; a 2 × CD-ROM (4 × recommended); 8 MB of RAM (16 MB recommended); and a mouse or drawing equivalent. Marine and aeronautical charts are also available. For a list of states and updates see the manufacturers website at <www.maptech.com>

Reader Enquiry No. 105

ArcData online

From Esri and Horizons Technology

Internet delivery of topographic raster maps

These companies have teamed up to deliver topographic raster maps over the Internet,



Raster any domain with ArcData Online.

using HTI's Internet map server and Esri's spatial database engine technology. By entering a street address or city name, visitors to Esri's innovative ArcData Online website can retrieve the topographic map of their area. This joint venture is designed to make high-quality, affordable topographical maps available to Esri customers for use with their geographic information systems (GIS) projects. Maps Raster is a series of full-colour, georeferenced United States Geological Survey (USGS) topographic raster maps, which can be obtained both on CD and over the Internet. Data are available in 1:24,000, 1:100,000 and 1:250,000 scales. Free viewing, extraction and reprojection software allows the user to scroll seamlessly across paper map boundaries, select an area, convert the map to a TIF file, reproject it and import it directly into raster-capable software.

Reader Enquiry No. 106

Hydronium and hydroxide

Turtle

From Hanna Instruments

Software that turns a PC into a pH meter

This Windows-based software package can turn a desk-top personal computer into a real-time pH meter for about the same price as a conventional pH electrode, says Hanna. The pH 'Turtle' software package provides a large, clear display that can be viewed on a PC monitor. This benefit could be particularly useful for educational purposes, allowing larger groups of people to see the information on screen. The package also allows users to plot the results and add notes, such as "added 5 ml of nitric acid at 10:03 a.m.". The software is not tied into just one computer; because it is multicensable, it can be loaded onto as many on-site computers as needed. The package provides data in tabular or graphic format with the facility to visualize pH and compensate for temperature from the keyboard. It includes high and low alarm settings with an audible warning. The Turtle software is compatible with Windows 3.1 and Windows95, has a pH range of 0.0–14.0, an accuracy of ± 0.2 pH, a resolution of 0.1 pH and can operate within a temperature range of 0–50 °C.

Reader Enquiry No. 107

SensorLink

From Orion

A new pH electrode measurement system that connects pH and ion-selective electrodes directly to your laptop or desktop PC

The manufacturer offers systems that allowing measurement of pH and/or concentration, for ion-selective electrode measurements, with temperature and voltage being displayed simultaneously. Conductivity, salinity, TDS and dissolved oxygen versions will be released at a later date. The software is said to make measurement and data collection very simple and offers improved measuring features, real-time graphs, data-logging and step-by-step user instructions. There are both Windows and a pH Palmtop (MS-DOS) versions available. The PCM500 pH for Windows system allows unlimited calibration points, provides drift correction and displays pH, mV and temperature simultaneously. The PCM700 pH/ISE/ORP for Windows extends measurement to additional direct read and incremental techniques for ion measurement.

Reader Enquiry No. 108

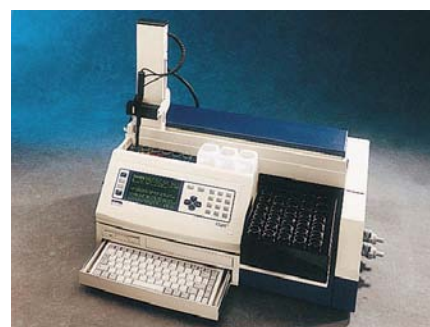
GLpH autosampler

From Sirius

Automates all aspects of pH measurement including calibration, electrode maintenance and all good laboratory practice reporting

GLpH indicates when buffers need changing and is designed for use by unskilled operators. It offers the option of total automation with a 32- or 50-place autosampler, and now has an optional autosampler to provide quality-control protocols. A sample run is organized so that dirty probes never pass over an unmeasured sample, and a different method may be selected for each sample, ensuring precision for each measurement. A 50-measurement run may now be taken in around five minutes. Furthermore, 'rush' samples may be introduced while using the autosampler, which has a tray that may also be connected to a water circulator to measure pH at stable temperatures. Calibration is performed automatically at user-programmed intervals and the system will store up to 20 pre-programmed methods. Electrodes are cleaned automatically between samples and the system regularly checks and logs electrode mV in pH 7 buffer, warning of deviations. Results or other data cannot be changed after collection, and there is a restricted access 'supervisor mode', which prevents tampering with methods. An audit trail log is automatically printed after maintenance operations. User names can be tagged with PIN numbers, and sample identification, results and calibration data may be stored in memory, or transferred directly to a PC or LIMS.

Reader Enquiry No. 109



pHantastic automation with the GLpH system.

pH540

From WTW Measurement Systems

A pH meter with full documentation of calibration and measurement for GLP and DIN compliance

Key features include real-time clock, automatic calibration, storage of 100 pairs of measurement data, display of data and time and an RS 232 interface.

Reader Enquiry No. 110

Glass-free pH sensor

From Select Systems

A new pH meter that uses a non-glass, stainless steel sensor

The probe stores dry and requires no maintenance. It should be useful in applications where safety is paramount and in environmental applications where accidental damage can easily occur. Owing to its shock-resistant rubber holster, this ultra-rugged system can withstand a 2-m drop onto concrete. Temperature compensation is automatic or manual, and up to three buffers are recognized automatically; there is also a choice between one calibration point or two. Depending on whether speed or accuracy is most important, the display resolution can be adjusted from 0.01 to 0.1 pH. Additional probes for measurements of flat surfaces, small samples and regular piercing of tissue samples are also available.

Reader Enquiry No. 111

What's on your microplate

BioPlate

From Polyfiltronics

A range of biochemically active microplates for solid-phase extraction (SPE) applications

SPE microplates come preloaded with a wide variety of sorbents with loadings per well ranging from 50–500 mg. Available formats include 96-well microplates with either an 800- μ l or 2,000- μ l well capacity and 24-well plates with 10-ml well capacity. All these plates have the standard footprint that fits most centrifuges, vacuum manifolds and liquid handlers (all are assembled with drip directors). In addition, Polyfiltronics has developed a number of acces-

sories for combinatorial chemistry, including the CombiClamp for sealing the top and bottom of BioPlates and vacuum manifolds. The self-contained Hi-Top system incorporates gas pressure beneath the plate to lift reagents off filter media during the reaction and subsequent cleavage step.

Reader Enquiry No. 112

Deepwell microplates

From Hamilton

Deepwell microplates constructed of polystyrene in a 96-well microplate format

These plates have a 1.2-ml capacity and should prove useful for archiving samples, PCR pooling, cell suspensions, large dilutions and mother-to-daughter plate pipetting. Deepwell microplates are available in cases of 32. In addition, an optional plate cover helps to ensure that each plate well is sealed, thus reducing cross contamination.

Reader Enquiry No. 113

Respair masks

From Sentinel Laboratories

An range of disposable respirators for comfortable, yet effective, protection against dust, mist, fume and fibre hazards

The Respair range of masks is disposable so there is no need for expensive maintenance or filters. This also means more hygienic protection because each mask has only one wearer. Lightly woven Technostat materials are said to make the masks fit closely and comfortably. The range comprises of six types of masks, including three 'valved' models. These masks, which are approved from FFP1 (S) to FFP3 (S), have the benefit of an exhalation valve that permits improved air-flow and extremely low breathing resistance. This valve also helps prevent safety spectacles from fogging up. Respair masks have an ergonomic, fold-flat design so they fit into a pocket when not being worn. This prevents contamination entering the mask, making them more hygienic to use after re-donning, says Sentinel.

Reader Enquiry No. 114

Coda

From BioRad

An automated microplate analyser that can increase enzyme immunoassay (EIA) efficiency

Running up to 270 samples at a time and carrying out five different assays simultaneously, this fully integrated microplate analysis system can handle sample dispensing, reagent addition, plate shaking, washing, incubating, reading and data reduction, all with walk-away operation. This system is said to be compatible with any microplate format. Moreover, it has programmable protocol steps, which enable each assay to be customized and saved (an editing function

allows methods to be changed at any time). Using Windows-based software with screen icon commands, no specialist training is required, and data can be seamlessly exported into most LIMS or quality-control systems. A broad range of EIA test kits is available, including those for tumour markers, infectious diseases, thyroid, fertility and steroid testing, and autoimmune screening.

Reader Enquiry No. 115

Microlute

From Porvair Sciences

A solid-phase extraction (SPE) microplate that has had its height reduced and is capable of processing 96 samples

Microlute has retained its 2-ml capacity per well but with a significantly reduced height, rendering volume extenders nugatory to requirements. This is said to enhance the ease of automation and the microplate is now compatible with all liquid handling and robotic devices. Each well has an individual drain spout, improving sample transfer and reducing crossover contamination. According to the manufacturer, this product improves SPE sample preparation times by up to 80 per cent. Porvair Sciences can pack a Microlute development plate with different sorbents from different manufacturers, as well as providing a mixture of sorbents in the same plate, which enables different separations to be carried out simultaneously.

Reader Enquiry No. 116

Odds 'n' ends

Millenium ULT freezers

From Jencons

A range of -80°C freezers

These freezers are available in both upright and chest formats, and are designed for durability and efficiency with a 90 per cent recyclable construction, CFC-free refrigerant and insulation. They are also backed by a 5-year parts and labour warranty.

Reader Enquiry No. 117

Syringe tip filters

From Corning Costar

Designed for the sterilization, clarification or pre-filtration of small sample volumes

The conventional round design comes in diameter sizes of 3 mm to 50 mm. The other style of filter, marketed under the brand name μStar , provides higher working pressures and higher working volumes while minimizing fluid retention when compared to traditional designs.

Reader Enquiry No. 118

Spindrive

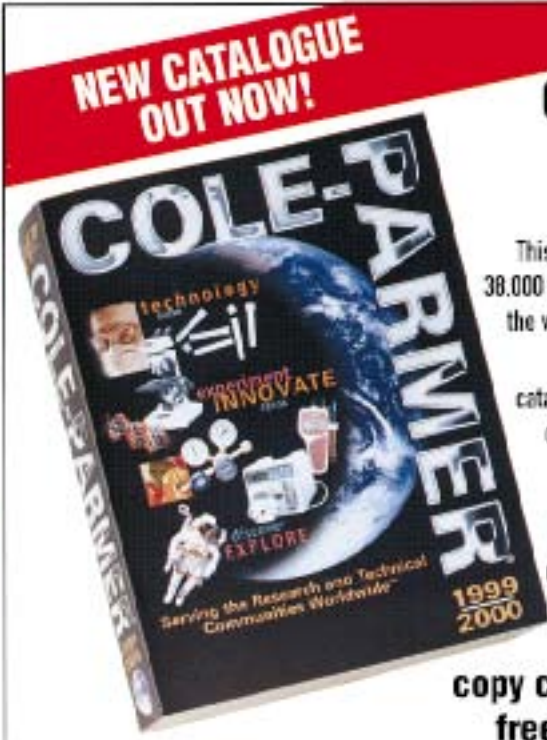
From Bel-Art

Converts a standard magnetic stirrer into a vortex mixer

This attachment uses a new vibrating platform, placed on top of the stirrer, to convert circular motion into a vortex with a diameter

ADVERTISEMENT

NEW CATALOGUE OUT NOW!



**COLE-PARMER
1999-2000
CATALOGUE**

This priced catalogue contains over 38,000 technical products from all over the world on 1300 full-colour pages.

Not only is it a comprehensive catalogue for purchasers and users of Laboratory equipment, it is an invaluable reference manual containing Technical Data and Chemical Resistance Charts for use by scientists and engineers.

**For your FREE
copy call Cole-Parmer on
free call 0500 345 300**

READER ENQUIRY NO. 45



Top spin: from magnetic stirrer to vortex mixer.

of 1.52 mm. This device is said to save space and money, and when use of the unit is complete, the user simply lifts the platform off the magnetic stirrer and stores it. The platform will mix up to four microtitre plates or two culture flasks at once. It features a non-skid surface and has a surface area of 225 mm².

Reader Enquiry No. 119

S7X detergent

From ICN

An environmentally friendly form of the well established laboratory detergent

The 7X PF (ES 7X in USA) detergent is a

concentrated liquid detergent made up of anionic surface-active agents and special solvents. It is free of phosphates and is biodegradable by normal environmental action. The detergent is immediately and completely soluble at any concentration, it contains strong sequestering reagents, disperses solids in suspension and drains from glassware in a few seconds, says ICN. The neutral pH will not induce etching of delicate glassware. Moreover, it is suitable for glass or plastic, test tubes, Petri dishes, flasks cylinders, pipettes, slide covers, and the like.

Reader Enquiry No. 120

Equipment online

Equipment search engine

From SciQuest.

A search engine for scientific products and services that is available free

As of January, scientific products and services from 11,788 vendors are listed on this web site. The database contains more than 175,000 product descriptions. It allows visitors to search, identify and communicate directly with vendors of their choice using the SciMail e-mail system. In 1998, SciQuest

is looking to increase the number of vendors in the database to 20,000 and to provide 500,000–700,000 product descriptions, all searchable by keyword at www.sciquest.com.

Reader Enquiry No. 121

Online equipment auction

From Internet Auctioneers Int. (IAI)

A guide to purchasing used lab equipment

The guide can be downloaded from IAI's Web site at <http://www.goinggoing-sold.com> or it can be ordered from the company. The internet auction service specializes in used and refurbished laboratory equipment. Visitors to the site can access information on available equipment and dates for the scheduled auction of each piece of equipment. Interested buyers submit their bids over the Internet during the specified open bidding time period. (IAI is a bonded auctioneer.)

Reader Enquiry No. 122

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENTS

Strike the Right Balance...

A range of Balances that deliver outstanding accuracy and reliability.

The HM, HF and HR series are designed for professionals who require the very best in laboratory weighing technology.

The Analyticals

HM Series

HF Series

HR Series

For more information please call
UK +44 (0)1235 550420 or
Germany +49 (0)69 5071017

<http://www2.aandd.co.jp>
e-mail: eurosales@aandd.co.uk

A&D INSTRUMENTS
...Clearly a Better Value

READER ENQUIRY NO. 46

Detect residual DNA with Wako's

DNA Extractor Kit

or Isolate Genomic DNA with

DNA Extractor WB Kit

Both use the Nal Method
No Phenol No Chloroform

Wako BioProducts
Wako Chemicals USA, Inc.
1600 Bellwood Road
Richmond, VA 23237
Telephone: (804) 271-7677
Facsimile: (804) 271-7791
(800) 992-9256
bioproducts@wakousa.com

READER ENQUIRY NO. 39

The publication you are reading is protected by copyright law. If you take photocopies from books, magazines or periodicals you should be licensed with the Copyright Licensing Agency (CLA). Make sure you are protected by a photocopying licence.

*The Copyright Licensing Agency Limited,
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931
Fax 0171 436 3986*